



Genome Structure of Mycobacteriophage D29: Implications for Phage Evolution

Michael E. Ford, Gary J. Sarkis, Aimee E. Belanger, Roger W. Hendrix
and Graham F. Hatfull*

The Pittsburgh Bacteriophage Institute and Department of Biological Sciences, University of Pittsburgh, Pittsburgh PA 15260, USA

Mycobacteriophage D29 is a lytic phage that infects both fast and slow-growing mycobacterial species. The complete genome sequence of D29 reveals that it is a close relative of the temperate mycobacteriophage L5, whose sequence has been described previously. The overall organization of the D29 genome is similar to that of L5, although a 3.6 kb deletion removing the repressor gene accounts for the inability of D29 to form lysogens. Comparison of the two genomes shows that they are punctuated by a large number of insertions, deletions, and substitutions of genes, consistent with the genetic mosaicism of lambdoid phages.

© 1998 Academic Press Limited

*Corresponding author

Keywords: mycobacteriophage; bacteriophage evolution; DNA sequence; genome organization; lysogeny

Introduction

The resurgence of tuberculosis in the United States and the prevalence of drug-resistant strains of *Mycobacterium tuberculosis* present considerable public health concerns (Bloom & Murray, 1992). Unfortunately, the genetics of the mycobacteria are not well understood although significant advances have been made in recent years (Hatfull *et al.*, 1994). The study of mycobacteriophages has proven particularly useful in the development of mycobacterial genetics. Mycobacteriophages are useful systems for characterizing their hosts and have been helpful in establishing mycobacterial genetic systems (Hatfull & Jacobs, 1994).

Mycobacteriophage L5 is the best characterized of the mycobacteriophages (Hatfull, 1994). L5 is a temperate phage isolated from a strain of *Mycobacterium smegmatis* by Doke (1960) that forms stable lysogens in which the phage genome is integrated into the bacterial attachment site, *attB* (Snapper *et al.*, 1988; Lee *et al.*, 1991). L5 also infects slow-growing mycobacteria such as *Mycobacterium bovis* bacille Calmette-Guérin (BCG), although rather specific conditions are required for efficient infection (Fullner & Hatfull, 1997).

The complete DNA sequence of the L5 genome has been determined (Hatfull & Sarkis, 1993). The genome of L5 is 52,297 bp long, composed of 85 protein-coding and three tRNA genes (Hatfull & Sarkis, 1993), and terminated in cohesive ends

(Oyaski & Hatfull, 1992). Although temperate, L5 encodes a DNA polymerase, tRNAs, and ribonucleotide reductase, features commonly associated with the lytic phages such as T4 (Brida & Abelson, 1985). Interestingly, L5 shares another attribute with lytic phages, the ability to inhibit host gene expression during lytic growth (Hatfull & Sarkis, 1993). While the identities of many L5 genes remain unknown, the genes encoding a number of the proteins which comprise the phage virion were revealed by amino-terminal sequences (Hatfull & Sarkis, 1993). Additionally, functional studies showed that the product of gene 33 (gp33) is the phage integrase (Lee *et al.*, 1991), and gp71 is the phage repressor responsible for maintenance of lysogeny and superinfection immunity (Donnelly-Wu *et al.*, 1993).

The protein encoded by gene 71 has been shown to bind to DNA and its binding site has been defined. Many such sites are scattered throughout the L5 genome, and low levels of transcription in the L5 prophage may be maintained *via* the binding of gp71 to these sites (Brown *et al.*, 1997). In addition, several L5 promoters have been located and studied, including those involved in modulating early growth and repressor transcription (Nesbit *et al.*, 1995). The repressor protein binds to an operator site overlapping the early lytic promoter P_{left} , thereby maintaining stable L5 lysogeny (Nesbit *et al.*, 1995; Brown *et al.*, 1997). Additionally, gp71 regulates its own synthesis by binding to a relatively weak binding site upstream of gene 71, modulating transcription from three promoters

Abbreviations used: PFU, plaque-forming units.

(P1, P2, and P3) in the gene 71-72 intergenic region (Nesbit *et al.*, 1995). Little is known about the location of signals responsible for late lytic gene expression.

Studies of mycobacteriophages have benefited mycobacterial genetics in a number of ways (Hatfull, 1994). First, integration-proficient vectors containing the L5 *int* gene and attachment site (*attP*) have been constructed which efficiently transform *M. smegmatis* and the vaccine strain *M. bovis* BCG (Lee *et al.*, 1991). These plasmids have been used to identify virulence genes (Pascopella *et al.*, 1994) and have proven useful in the production of recombinant BCG vaccines (Stover *et al.*, 1991). In addition, the L5 repressor gene has been successfully employed as a selectable marker for plasmid maintenance without the use of antibiotic resistance markers (Donnelly-Wu *et al.*, 1993). Recombinant L5 phages containing a copy of the firefly gene (*FFlux*) have been created and shown to be very effective for the rapid determination of antibiotic susceptibility patterns of mycobacterial isolates (Sarkis *et al.*, 1995). Similar reporter phages based on TM4 and D29 have also been described (Jacobs *et al.*, 1993; Pearson *et al.*, 1996). Recently, it has been shown that mycobacteriophages hold considerable promise as tools for transposon delivery in mycobacteria (Bardarov *et al.*, 1997).

In addition to the clues that were provided about its life cycle, several other interesting observations were brought to light by the sequencing of L5. Although L5 shares absolutely no discernible sequence similarity, at either the DNA or the amino acid levels, with lambdoid phages, some elements of the arrangement of its head and tail genes are reminiscent of phages such as lambda and P22 (Casjens *et al.*, 1992). Additionally, L5 apparently forms covalent crosslinks among all copies of its major head subunit during formation of the viral capsid (Hatfull & Sarkis, 1993; Hatfull & Jacobs, 1994), something previously documented only in the lambdoid coliphage HK97 (Popa *et al.*, 1991; Duda *et al.*, 1995). Finally, a programmed translational frameshift that has been shown to take place in lambda between genes G and T of the tail gene cluster (Levin *et al.*, 1993) also appears to occur in the analogous regions of HK97 and L5 (Casjens *et al.*, 1992; Hatfull & Jacobs, 1994), although these three phages show no sequence similarity in this area of their genomes.

Studies of members of the lambdoid group of bacteriophages have noted that the genomes of these phages appear to be genetic mosaics. In other words, highly similar segments of any two genomes are often separated by sharp transitions from adjacent segments that match each other at a different level of similarity or not at all. Casjens *et al.* (1992) take the appearance of such distinct boundaries between genome segments as evidence of multiple recombination events within the evolutionary histories of the phages that display them. These recombination events probably can occur anywhere within the phage genome, but only

recombination events that do not affect the viability of the resulting viruses survive. Therefore, such transition sites between matching and non-matching sequences of two phage genomes are witnessed most often at gene boundaries or at the boundaries of functional clusters of genes. However, additional examples are occasionally found within genes, at domain boundaries (Casjens *et al.*, 1992).

Mycobacteriophage evolution could be studied by comparing the genome of L5 to that of another closely related mycobacteriophage. Therefore, we determined the complete DNA sequence of mycobacteriophage D29. D29 was first isolated from soil, and shown to be active against *Mycobacterium tuberculosis* (Froman *et al.*, 1954). The plaques produced on the tubercle bacillus were clear, indicating that D29 is a lytic phage (i.e. incapable of lysogeny). Studies of D29 suggest that it is a member of the "L5-like" family of mycobacteriophages. Although lytic, D29 is subject to superinfection immunity by L5; in other words, it is incapable of infecting an L5 lysogen of *M. smegmatis*. Only the product of L5 gene 71 is required to prevent infection of *M. smegmatis* by D29, demonstrating that this is true immunity rather than exclusion (Donnelly-Wu *et al.*, 1993). In addition to the shared immunity of L5 and D29, these phages appear to have a common mode of entry in *M. smegmatis*. In particular, overexpression of the *M. smegmatis* *mpr* gene confers resistance to both L5 and D29, but not to other mycobacteriophages (Barsom & Hatfull, 1996). The identity of the receptor used for adsorption of these phages is not clear, although pyruvylated, glycosylated acyltrehaloses have been implicated in the infection of *M. smegmatis* by D29 (Besra *et al.*, 1994). Additionally, the two phages have similar host ranges (our unpublished observations), though there are some differences in infection requirements (Fullner & Hatfull, 1997), and D29 has been reported to adsorb to *Mycobacterium leprae* (David *et al.*, 1984). Finally, DNA hybridization studies showed that the D29 and L5 genomes are closely related (M. Donnelly-Wu & G.F.H., unpublished observations), although they show different patterns of fragments for many restriction enzymes (Lazraq *et al.*, 1989; Oyaski & Hatfull, 1992).

Here we describe the complete DNA sequence of the genome of mycobacteriophage D29, compare the genetic map obtained from it with that already determined for L5, and consider the consequences for bacteriophage evolution previously only intensively studied in the lambdoid family of phages. Some genetic explanations for observed phenotypic differences between L5 and D29 are given.

Results

Analysis of D29 particles

While the morphology of D29 particles was reported previously (Shafer *et al.*, 1977), no specific similarity to L5 (or any other mycobacteriophage)

was noted. We therefore assessed the morphologies of L5 and D29 by electron microscopy and found them to be virtually identical (Figure 1). We have also compared the virion proteins of D29 and L5 by SDS-PAGE of whole particles (Figure 2). The patterns are similar but not identical. In particular, we note that D29, like L5, has several protein species larger than 150 kDa (bands A, B, and C). In L5, this reflects the presence of extensive covalent crosslinking between subunits of the major capsid protein, gp17 (Hatfull & Sarkis, 1993). Presumably, the major head subunit of D29 is also extensively crosslinked. Indeed, amino-terminal sequencing of the D29 equivalent of band C revealed that it is the product of gene 17 (data not shown).

Many of the L5 proteins (including bands F, H, and J) are represented in similar sizes in D29, although proteins D and E are either absent or in low abundance in D29 (Figure 2). Additionally, a band corresponding to the L5 band I appears to be missing in D29. Instead, D29 has a protein band that migrates only slightly faster than band H. Finally, the intensity of band F is lower in D29 than in L5. In L5, band J corresponds to the major tail subunit protein gp23 (Hatfull & Sarkis, 1993). Sequencing of the amino-terminal end of the D29 equivalent of band J (data not shown) revealed that the first ten amino acids of this protein are identical to the first ten amino acids of L5 gp23, and also match the predicted N-terminal sequence of D29 gp23. Therefore, gp23 most likely constitutes the major tail subunit of D29.

Determination of the D29 genomic sequence

The complete DNA sequence of the D29 genome was determined using a shotgun sequencing approach. Phage D29 DNA fragments (1 to 3 kilobase pairs; kb) were ligated into appropriate vectors. Following transformation of *E. coli*, DNA was prepared from randomly chosen colonies and sequenced from both ends using standard dideoxy methods and an automated DNA sequencer.

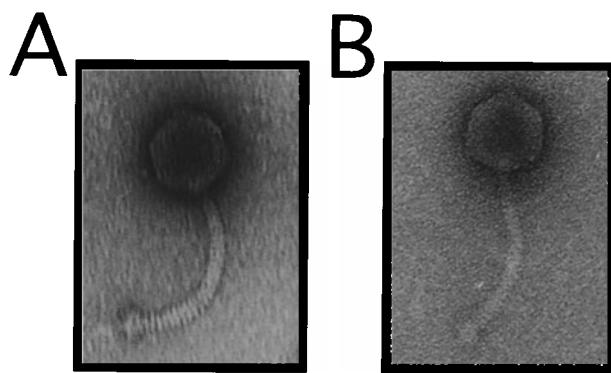


Figure 1. D29 and L5 virions. Electron micrographs of D29 (A) and L5 (B), taken at approximately the same magnification. Cesium chloride purified phage preparations were stained with uranyl formate.

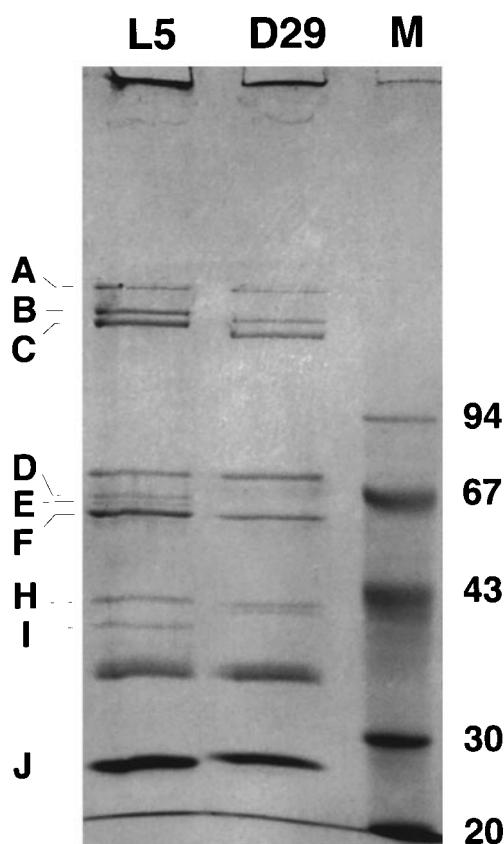


Figure 2. Analysis of the structural proteins of L5 and D29. SDS-PAGE comparison of the structural proteins that comprise intact L5 and D29 phage particles. The L5 proteins are given a letter corresponding to their identification by Hatfull & Sarkis (1993). The smeared band below band I is DNaseI, which was added to the preparation for removal of phage chromosomal DNA. The band above band D is BSA, a component of the DNaseI stock that was used.

Assembly of approximately 1100 such random DNA sequences indicated that D29 phage particles contain a linear double-stranded DNA molecule 49,136 base-pairs (bp) in length (GenBank accession no. AF022214). The average useful length of each sequence, after end-trimming and vector removal, was 348 bases and each base was sequenced on average 7.7 times. The G + C content of the D29 genome is 63.6%, a value consistent with that of the mycobacterial hosts (Clark-Curtiss, 1990), and virtually identical to the G + C content of L5 (63.2%; Hatfull & Sarkis, 1993). Sequencing directly from D29 DNA we found that the D29 chromosome contains cohesive termini identical to those found in L5 (5'-GGTCGGTTA-3' with a 3' extension; data not shown).

Assignments of probable D29 genes

D29 genes were identified using a previously constructed L5 codon usage table (Hatfull & Sarkis, 1993), as well as the programs of Staden

Table 1. Coordinates of D29 genes and calculated molecular masses (kDa) of predicted polypeptides

Gene	kDa	Frame	Start	Stop	Comments
Left arm					
1	30.3	2	401 (GUG)	1213 (UAG)	
2	28.8	1	1327 (AUG)	2106 (UAA)	
3	9.0	3	2106 (AUG)	2357 (UGA)	
4	11.1	2	2354 (GUG)	2650 (UAA)	
5	17.1	1	2686 (GUG)	3153 (UGA)	
6	34.8	3	3240 (AUG)	4208 (UAG)	
7	—	—	4218	4292	Minor tail protein
8	—	—	4333	4408	Asn-tRNA
9	—	—	4416	4488	Trp-tRNA
9.1	—	—	4523	4596	Gln-tRNA
9.2	—	—	4600	4680	Glu-tRNA
10	54.8	2	4706 (AUG)	6187 (UGA)	Tyr-tRNA
11	14.6	1	6184 (AUG)	6609 (UGA)	
12	28.5	3	6606 (AUG)	7370 (UGA)	
13	66.4	1	7384 (GUG)	9171 (UGA)	
14	53.5	3	9168 (AUG)	10,625 (UGA)	
15	31.2	2	10,655 (GUG)	11,482 (UAA)	
16	19.8	2	11,498 (GUG)	12,055 (UGA)	Scaffold protein?
17	33.9	3	12,096 (AUG)	13,052 (UGA)	Major head subunit
18	5.7	3	13,122 (AUG)	13,277 (UAA)	
19	14.1	3	13,281 (AUG)	13,655 (UGA)	
19.1	7.1	2	13,652 (AUG)	13,846 (UGA)	
20	13.8	1	13,846 (AUG)	14,214 (UGA)	
21	12.2	3	14,214 (AUG)	14,549 (UAA)	
22	15.4	3	14,559 (AUG)	14,978 (UAA)	
23	21.3	3	14,994 (AUG)	15,590 (UGA)	Major tail subunit
24	15.2	2	15,704 (AUG)	16,108 (UGA)	
25	11.4	1	16,222 (GUG)	16,527 (UAG)	
26	86.7	2	16,517 (AUG)	19,030 (UGA)	Minor tail subunit
27	38.7	1	19,051 (UUG)	20,061 (UGA)	Minor tail subunit
28	67.1	3	20,058 (AUG)	21,848 (UAA)	Minor tail subunit
29	17.0	1	21,865 (AUG)	22,308 (UGA)	
30	10.4	3	22,305 (GUG)	22,571 (UGA)	
31	65.6	2	22,568 (UUG)	24,421 (UAA)	
32	21.8	1	24,421 (AUG)	25,092 (UAA)	
32.1	7.6	3	25,101 (AUG)	25,304 (UGA)	
33	40.0	6	26,429 (GUG)	25,359 (UGA)	Integrase
Right arm					
34.1	15.0	3	26,874 (GUG)	27,311 (UGA)	Excise?
36	6.3	6	27,605 (AUG)	27,435 (UGA)	
36.1	13.5	6	27,992 (AUG)	27,606 (UGA)	Deoxycytidylate deaminase?
38	4.9	5	28,144 (AUG)	27,992 (UGA)	
39	14.8	4	28,527 (AUG)	28,141 (UGA)	
41	11.1	5	28,798 (AUG)	28,514 (UGA)	
41.1	7.6	6	28,982 (AUG)	28,782 (UGA)	
42	7.4	5	29,188 (AUG)	28,892 (UGA)	
43.1	5.4	4	29,328 (UUG)	29,188 (UGA)	
44	68.2	5	31,162 (AUG)	29,339 (UGA)	DNA Polymerase
44.1	12.2	6	31,505 (AUG)	31,170 (UGA)	
46	15.3	5	31,915 (AUG)	31,505 (UGA)	
48	26.5	4	32,832 (AUG)	32,125 (UGA)	
49	20.3	5	33,478 (UUG)	32,900 (UGA)	
50	77.2	4	35,571 (GUG)	33,490 (UGA)	Ribonucleotide reductase?
51	7.3	6	35,768 (GUG)	35,571 (UAG)	
52	7.1	5	35,950 (AUG)	35,765 (UGA)	
53	26.3	6	36,641 (UUG)	35,934 (UGA)	
53.1	6.0	5	36,790 (UUG)	36,638 (UGA)	
54	28.5	5	37,561 (GUG)	36,794 (UGA)	
55	17.3	6	38,009 (GUG)	37,554 (UAA)	
56	10.2	5	38,278 (AUG)	38,006 (UGA)	
57	17.2	4	38,736 (AUG)	38,278 (UGA)	
58	14.1	5	39,127 (UUG)	38,738 (UGA)	DNA primase?
59	17.6	4	39,585 (GUG)	39,124 (UGA)	T4 Exo VII?
59.1	5.0	5	39,715 (AUG)	39,572 (UAA)	
59.2	29.8	4	40,548 (AUG)	39,712 (UGA)	Haloperoxidase
61	14.0	6	40,922 (AUG)	40,545 (UGA)	
62	5.8	5	41,077 (GUG)	40,922 (UGA)	
63	8.9	4	41,310 (AUG)	41,074 (UGA)	
64	14.0	4	41,859 (AUG)	41,473 (UGA)	
65	25.1	5	42,589 (UUG)	41,891 (UGA)	
66	23.0	5	43,363 (GUG)	42,764 (UGA)	

Table 1—Continued

Gene	kDa	Frame	Start	Stop	Comments
66.1	5.2	4	43,500 (GUG)	43,360 (UGA)	
68	8.7	6	43,730 (AUG)	43,497 (UGA)	
68.1	9.4	4	44,013 (AUG)	43,771 (UGA)	
69	30.7	6	44,870 (GUG)	44,061 (UGA)	
70	15.7	5	45,286 (AUG)	44,867 (UGA)	
82/71	—	6/4	45,903 (AUG)	45,342 (UAA)	
82.1	5.1	6	46,031 (GUG)	45,900 (UGA)	
82.2	3.6	5	46,123 (AUG)	46,028 (UGA)	
84	6.9	4	46,314 (GUG)	46,120 (UGA)	
96	10.1	6	46,595 (AUG)	46,329 (UGA)	
87	6.2	5	46,762 (GUG)	46,598 (UGA)	
88	26.5	6	47,492 (AUG)	46,770 (UGA)	
89	11.6	6	47,939 (AUG)	47,628 (UGA)	

(1986) and GeneMark (Borodovsky & McIninch, 1993). The beginnings of protein-coding regions were identified as positions of a sharp rise in coding potential (based on codon usage) that coincided with a potential start codon and, in most cases, a good Shine-Delgarno ribosome binding site. Three potential translation initiation codons were identified: AUG (47 examples), GUG (22 examples), and UUG (eight examples). Initial evidence for the use of all three start codons was provided in the L5 genome sequence analysis (Hatfull & Sarkis, 1993) and further support is provided by the comparative analysis described below. The end of each coding region was defined by one or more of the standard termination codons. Using these methods a total of 77 putative protein-coding genes was identified. We also identified a cluster of five tRNA genes using the programs of Staden (1986). A list of the coordinates of the D29 genes is provided in Table 1.

Global organization of D29

The complete genome sequence of D29 supports the hypothesis that it is very closely related to L5. The DNA sequences of many of the genes are very similar, and although the D29 genome is slightly shorter than L5's (49,136 bp versus 52,297 bp) the genomes are generally colinear (Figure 3). We have therefore numbered the D29 genes in accord with the L5 nomenclature. Specifically, D29 genes whose products show homology to their L5 counterparts are given the same number as in L5 (e.g. D29 gene 33 is homologous to L5 gene 33). Where additional or non-homologous genes are present, a two-part number has been assigned; the first integer corresponds to the preceding gene (numerically) and is followed by a period and a second digit (e.g. the two D29 genes following gene 59 are numbered 59.1 and 59.2).

Although D29 is not a temperate phage, it does contain a phage attachment site (*attP*) that is very similar to L5's, including the common core (although one position at the extreme end is different such that only 42 bp are common to the *M. smegmatis attB* sequence; see Peña *et al.*, 1997). The D29 attachment site (coordinates of the common core are 26,634 to 26,676) is located somewhat

to the right of center of the genome and divides the genome into a left and right arm as described for L5 (Figure 3). The left arm genes are closely spaced and, with the exception of the phage integrase (gene 33), are transcribed rightwards. All of the right arm genes, with the exception of gene 34.1, are transcribed leftwards. The only significant non-coding regions lie adjacent to the left and right *cos* sites.

While the D29 and L5 genomes are closely related, the level of similarity is not the same across the genomes. In particular, the left arms can be aligned without introducing substantial gaps and are approximately 80% identical at the nucleic acid level. The right arms are considerably less closely related although the differences are not evenly distributed. Thus, in the right arm, regions of high similarity are punctuated by segments of unrelated DNA (see below).

Modifications to the L5 genome map

Although the genome sequences of L5 and D29 were determined and analyzed independently (i.e. the L5 analysis was not used to determine coding regions in D29) the maps are very similar, adding confidence to the accuracy of the sequence determinations and veracity of the analyses. In addition, the comparative analysis benefited the study of L5 in that it suggested that some modifications to the L5 map were needed. In particular, five small open reading frames (19.1, 43.1, 53.1, 68.1 and 89) were not previously assigned in L5, but since they are present and closely related in D29, they are included in the revised L5 map (Figure 3). In one instance, an L5 gene (34; immediately to the right of *attP* and transcribed leftwards) appears to have been improperly assigned, since the reading frame is not conserved in D29. However, a second reading frame (34.1) on the other strand is present in both phages and is a better candidate for a protein-coding gene even though the coding potential (judged by codon usage) is not particularly strong in either phage. The lengths of the two predicted gene products are identical (145 residues), although only the first 121 residues are similar (69% identity). Gene 34.1 is located in a region of the phages' chromosomes that may contain the

L5

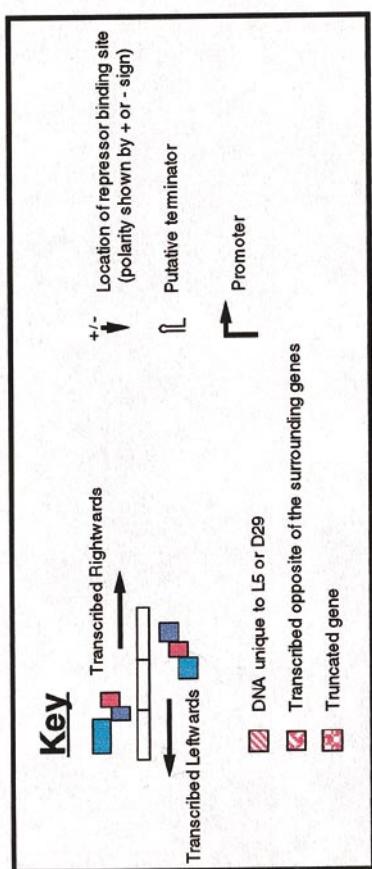
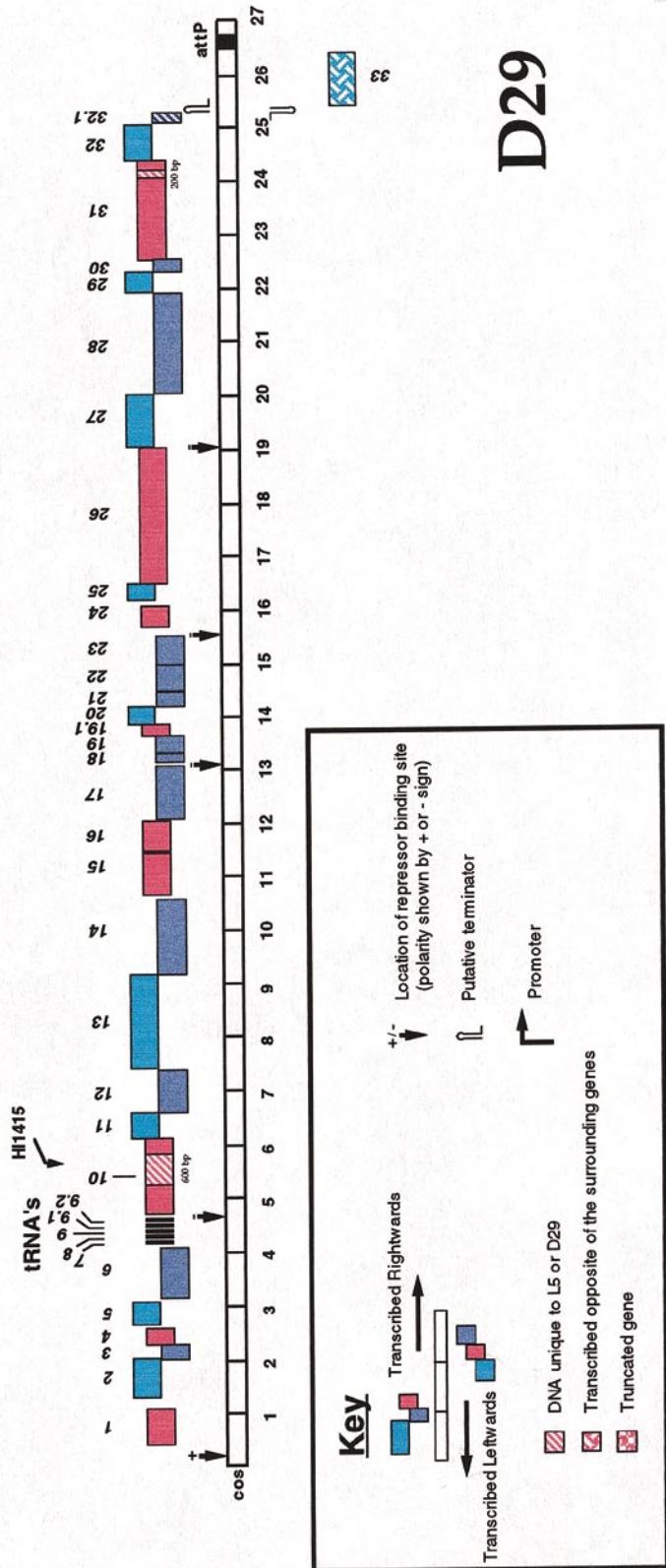
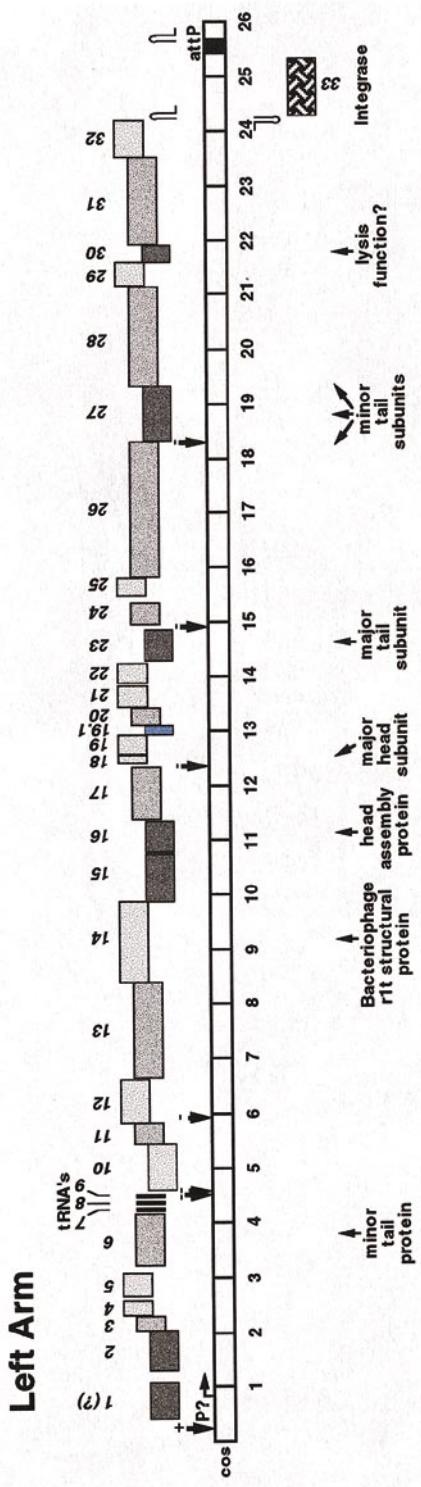
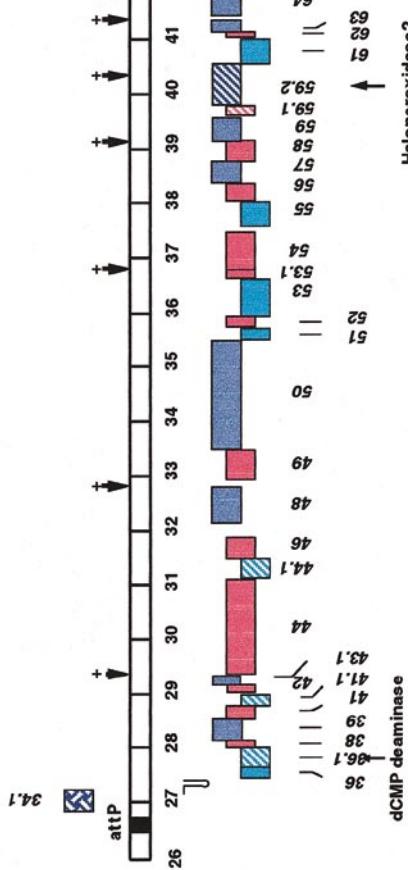
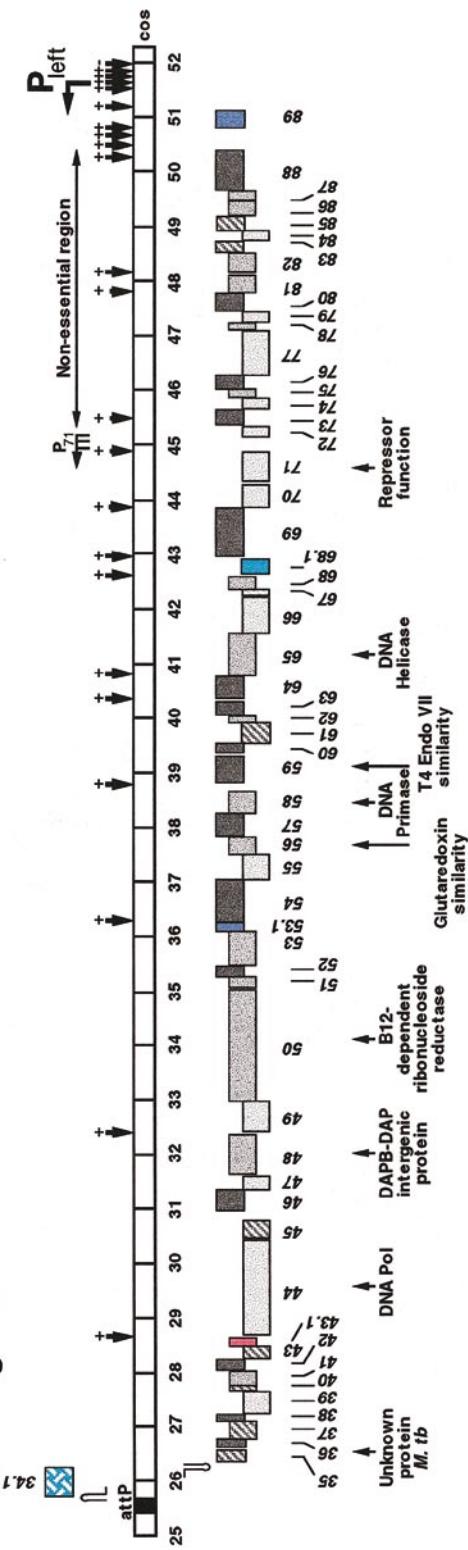


Figure 3 (legend opposite)

L5

Right Arm



D29

Figure 3. Genetic organization of D29 and L5. Comparison of the open reading frame maps of L5 and D29. The two phages' genomes are represented by white horizontal bars with vertical markings every 1000 bp. Open reading frames are depicted as horizontal boxes of three different shadings and heights to show which open reading frame is used. D29 genes and recently identified L5 genes are shown in color. For more information, refer to the key. The scheme used for naming the genes is described in the text.

Table 2. Previously unidentified L5 genes

Gene	kDa	Frame	Start	Stop
19.1	7.1	3	12,933 (GUG)	13,127 (UGA)
34.1	15.1	1	25,777 (GUG)	26,220 (UGA)
43.1	6.7	5	28,678 (UUG)	28,499 (UGA)
53.1	6.0	4	36,264 (UUG)	36,112 (UGA)
68.1	11.6	6	42,962 (UUG)	42,657 (UGA)
89	11.2	5	51,082 (AUG)	50,786 (UGA)

excisionase functions (Lee *et al.*, 1991). Finally, while a segment of DNA homologous to the coding sequence of L5 gene 47 is present in D29, there is a one base deletion (relative to L5) which changes the reading frame and produces a shorter putative gene product. We presume that either the L5 or D29 gene 47 product is non-functional, although we cannot rule out an error in one of the sequences. A list of the newly assigned L5 genes is given in Table 2. Finally, several database matches have been identified since the L5 sequence was reported and these are indicated in Figure 3.

Predicted gene functions

Each of the 77 predicted protein coding regions of the D29 genome was translated into its amino acid sequence and searched against the protein databases using the BLAST and FASTA programs of GCG (Genetics Computer Group, Madison, WI). The majority of these searches showed similarity only to L5 sequences already in the database, and in most cases this similarity was strong (average amino acid identity, 74%). Using such comparisons, we were able to find the L5 homolog for most of the predicted D29 gene products, including every L5 gene for which a function is known. Our analysis of the D29 genome does not confirm earlier reports that D29 may encode a lipase and an RNA polymerase (Jones & David, 1970, 1971).

We identified three database matches for D29 gene products for which there is no homolog in L5. One of these is an internal segment of gp10 (which is absent from L5) that matches HII1415, a hypothetical protein of unknown function from *Haemophilus influenzae* Rd (Fleischmann *et al.*, 1995). The remainder (genes 36.1 and 59.2) have reasonable similarities to deoxycytidylate deaminases and non-heme haloperoxidases, respectively. All three are discussed in more detail below.

Genetic basis of D29 lytic phenotype

Although D29 is a close relative of temperate L5 and has an *attP*-integration system, it does not form lysogens. The simplest explanation for D29's exclusively lytic lifestyle is a 3.6 kb deletion relative to L5 at the right end of its genome which removes part of the repressor gene (71) and several adjacent genes (Figure 4). Alignment of the genome sequences indicates that D29 has lost 3620 bp relative to L5 (between L5 coordinates 44,700 and 48,321; the junction of the flanking sequences in D29 occurs between coordinates 45,706 and 45,707)

probably through a simple deletion without substantial additional rearrangements (Figure 4A). However, the mechanism giving rise to this deletion is not obvious, since, at least within the L5 sequence, there is little sequence similarity between the flanking regions. Deletion derivatives of L5 which have lost all or part of the repressor gene (71) and are incapable of forming lysogens have also been isolated (Donnelly-Wu *et al.*, 1993).

By comparing the L5 and D29 coding regions at the point of deletion, it is apparent that D29 has lost about half of both genes 71 and 82 and presumably all of the intervening genes (Figure 4B). In L5, gene 71 encodes the phage repressor, and is required for the maintenance of lysogeny and is sufficient to confer superinfection immunity (Donnelly-Wu *et al.*, 1993). The observation that D29 does not have an intact copy of gene 71 thus completely explains its lytic properties. In addition, because D29 presumably has lost L5 genes 71 to 82, this region is most likely not essential for phage growth, an observation consistent with previous reports that the entire region from 71 to 88 is non-essential (Sarkis *et al.*, 1995). Little is known about the functions of any of the genes in this area of the L5 genome, although a gene in the 72 to 82 interval is potentially involved in lysogenic establishment (Sarkis *et al.*, 1995).

While the 3.6 kb deletion adequately accounts for the lytic phenotype of D29, the appearance of this deletion could have been an artifact of faulty subclone assembly during sequence compilation. To confirm that the deletion was indeed a feature of its genome, D29 was sequenced directly using primers on either side of the putative deletion and close to the end points. This analysis confirmed that D29 carries the deletion described above.

Lysogenization of D29 in a gp71⁺ strain of *M. smegmatis*

As stated previously, D29 contains an intact *attP*-integration system much like the one found in L5. In addition, the majority of the repressor protein binding sites found in L5 are present also in D29 (these will be discussed in more detail later). Therefore, it seems reasonable to suppose that D29 could lysogenize *M. smegmatis* if the L5 repressor, gp71, were present. To test this, we infected a strain of *M. smegmatis* containing the L5 repressor gene (*mc*²155(pMD132)) growing on solid medium with a recombinant D29 phage (phBD8) that carries the firefly luciferase gene (Pearson *et al.*, 1996). Following incubation, cells were picked from the infected area and plated for single colonies (Although the L5 repressor confers immunity to D29, cell lysis results when high titers of phage are spotted onto lawns of cells containing the repressor.) Approximately 50% of the colonies were found to produce light (Figure 5A), indicating that they carry the D29:FFlux phage. A similar proportion of light-producing colonies was recovered after infection with phGS26, a clear-plaque FFLux

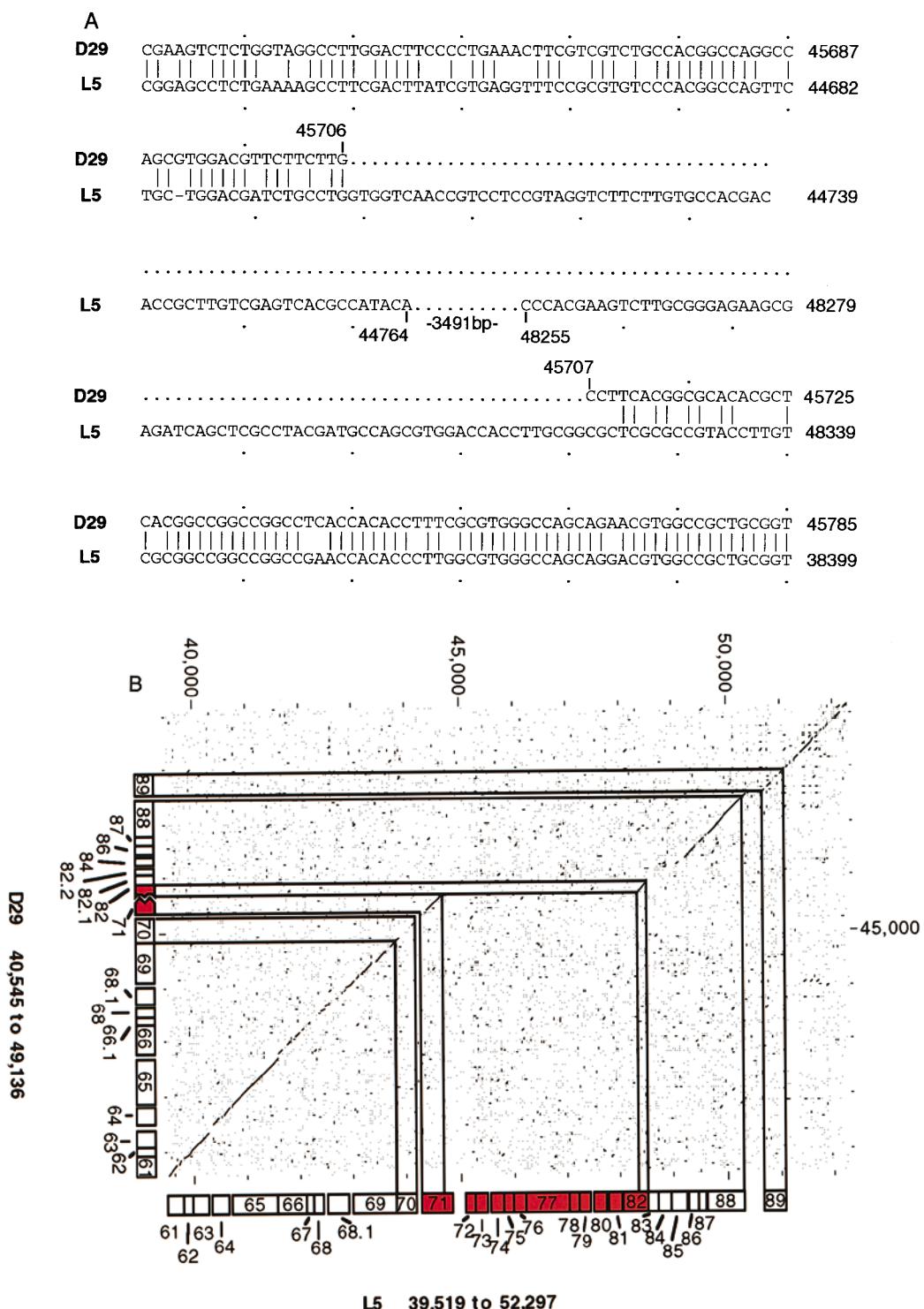


Figure 4. Deletion of the immunity region of D29. A, Bestfit DNA alignment of L5 and D29 showing the endpoints of the deletion of 3.6 kb of DNA from D29 which removes a portion of the repressor gene (71) and several others in their entirety. B, Diagon plot comparing the immunity and surrounding regions of the L5 genome with the corresponding area of the D29 chromosome. Open reading frames have been drawn on for reference, in addition to vertical and horizontal lines which emphasize important areas. Genes 71 and 82 of D29 are separated by a broken boundary to indicate that they are truncated forms of their L5 counterparts. Genes affected by the D29 deletion are shown in color for both phages. Note the low similarity of some of the genes surrounding the immunity regions of L5 and D29 (e.g. 70 and 84), as well as the large intergenic space between L5 genes 88 and 89 that is not present in D29.

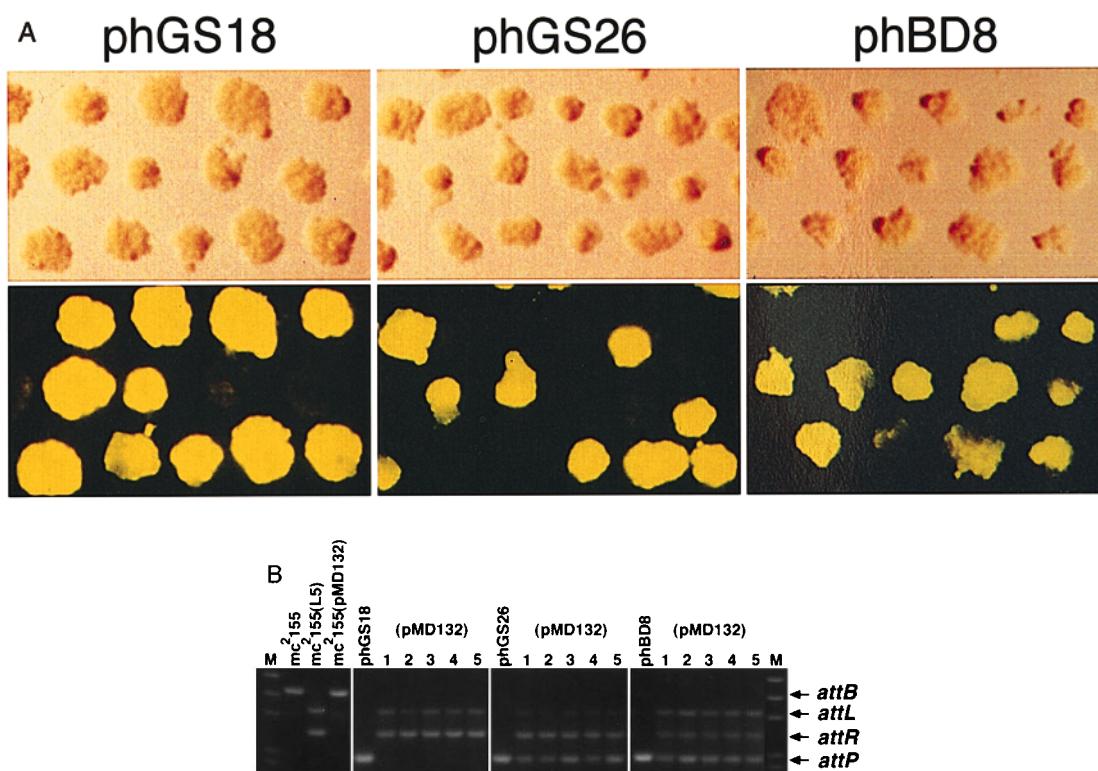


Figure 5. D29 lysogenization of *M. smegmatis* carrying L5 gene 71. A, Isolated colonies were recovered from *M. smegmatis* cells carrying the L5 repressor gene (*mc*²155(pMD132)) that had been infected with a temperate L5 luciferase reporter phage (phGS18), a clear plaque L5 luciferase reporter phage (phGS26) or a D29 luciferase reporter phage (phBD8). Colonies were picked onto fresh plates, incubated and photographed under incident light, or in the dark following addition of a luciferin solution. Approximately 50% of the cells recovered from phGS26 or phBD8 infection are light-producing lysogens. B, Five of each of the phGS18, phGS26 and phBD8 light-producing lysogens were used for PCR analysis, using a mix of four primers that amplify either the chromosomal *attB* site, the phage attachment sites of L5 and D29 (*attP*) or the attachment junctions *attL* and *attR*.

derivative of L5 (Figure 5A). Both the phBD8 and phGS26-infected light producing colonies were analyzed by PCR amplification of the bacterial and phage attachment sites, and it was found that all of them contained *attR* and *attL* attachment junctions. These results demonstrated that D29 (in the case of phBD8) or L5 (in the case of phGS26) had integrated into the *M. smegmatis* genome at the *attB* site (Figure 5B). However, *attP* DNA was amplified in addition to the *attL* and *attR* junctions, suggesting that these lysogens are somewhat less stable than lysogens formed by a temperate L5 FFlux phage such as phGS18 (see Figure 5B).

Genomic differences: genetic mosaicism of the L5-like phages

While the D29 and L5 genomes are generally colinear, significant differences between the two genomes reflect a mosaic-like genome arrangement similar to members of the lambdoid bacteriophage family. Several varieties of genomic differences are present, including insertions, deletions, and substitutions. The specific nature of these discontinuities is of considerable interest, since it sheds light on the mechanisms of phage evolution. Examples of each will be discussed.

Insertions and deletions

In general, where one genome contains a DNA segment that the other does not possess it is difficult to know whether this represents an insertion of DNA into one genome or loss of DNA from the other. Thus all of these types of differences will be discussed together. First, there are several examples where D29 contains DNA segments not present in L5. In three such cases, these represent additions of DNA in D29 between the homologs of previously assigned L5 genes. One example can be found between genes 9 (tRNA^{Gln}) and 10 (unknown function) where D29 has a segment of approximately 190 bp not present in L5 (Figure 6A). This region codes for two additional tRNA genes (tRNA^{Glu} and tRNA^{Tyr}) as shown in Figure 6B. In this case, we favor the explanation that L5 has lost two tRNA genes as a result of a deletion. However, it is also noteworthy that the spacing between the tRNA genes that the two phages share is different, indicating the occurrence of other deletion or insertion events. In a second example, D29 has a DNA segment inserted between genes 32 and 33 which contains a newly assigned gene, 32.1. Since 32 is the last gene in the late operon of L5 and gene 33 is transcribed in the leftwards direction, 32.1 is the last

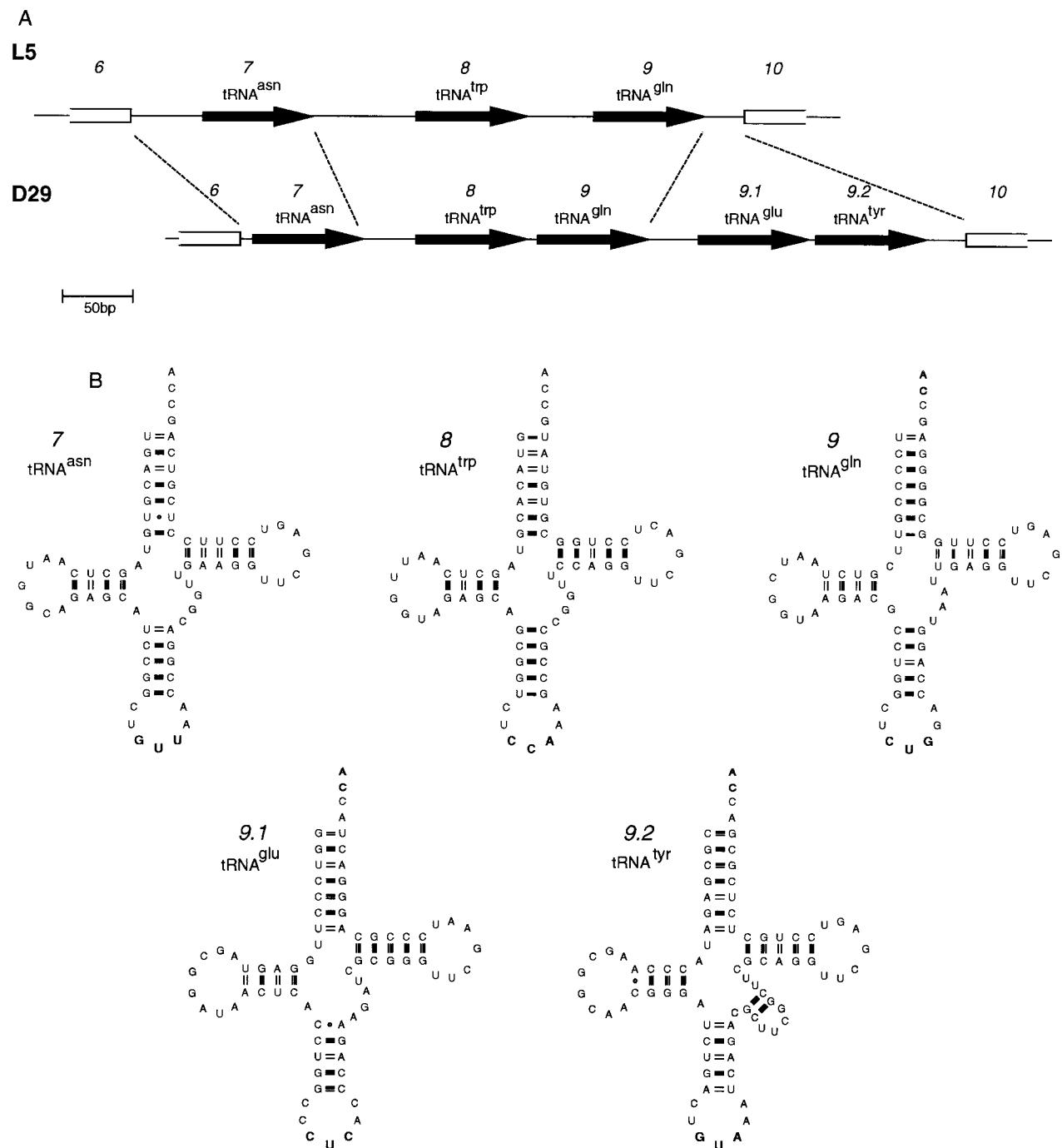


Figure 6. The tRNAs of L5 and D29. A, Schematic alignment of the tRNA cassettes of L5 and D29. Note that, in addition to possessing two tRNA-like genes that L5 does not, D29 displays a different pattern of intervening DNA between the tRNA-like genes it shares with L5. B, Predicted secondary structures of the tRNA molecules encoded by D29. Anticodons are indicated in bold print, as are the nucleotides CA at the 3' ends of the products of genes 9, 9.1, and 9.2, which are presumed to be added post-transcriptionally.

gene in the late operon of D29. Again, we cannot rule out acquisition of this gene by D29 but note that precise deletion of a 212 bp D29-like precursor could have given rise to the L5 organization, retaining not only genes 32 and 33 but also the hairpin-loop terminator-like structure located between these genes (see Figure 7A and B). However, it is interesting to note that the D29 gene

32 is only 58% identical to the L5 allele, indicating that something more substantial than a simple deletion may have taken place (the average gene identity is 75%). D29 gene 41.1 is also not present in L5, and the L5 genes 35, 40, and 85 are not present in D29. Again, their absence from D29 can be accounted for by relatively simple deletion events (Figure 7C).

Two D29 genes are substantially larger than their homologs, through the presence of additional DNA within the coding region. One of these is gene 31, which is about 200 bp longer in D29 than in L5, largely resulting from the presence of additional DNA within the coding region close to the 3' end of the gene. However, there are other differences (the DNA sequences immediately 5' to the discontinuity are quite different) between these genes and the relationship cannot be simply described as an insertion or deletion of a specific DNA segment. The second example is gene 10, where the D29 homolog is approximately 600 bp larger than its L5 counterpart as a result of additional DNA near to the middle of the gene. The portion that is unique to D29 matches an open reading frame of similar length and unknown function in the *H. influenzae* genome (HI1415; Fleischmann *et al.*,

1995). The relationship between these genes is the subject of a separate study (R.W.H., M.F. & G.F.H., unpublished). It is noteworthy that the sections of the D29 gene product that flank the novel internal segment match the corresponding sections of L5 gp10 at quite different levels. The N-terminal section shares 79.2% identity with the corresponding section of the L5 protein while the C-terminal portions are only 49.6% identical. Perhaps the different segments of this gene represent distinct functional domains which are under different selective pressures.

Gene substitutions

There are several cases of apparent gene substitutions, in which an established L5 gene is not present in D29, but rather has been replaced by a novel gene or genes. One example is the gene 36 to

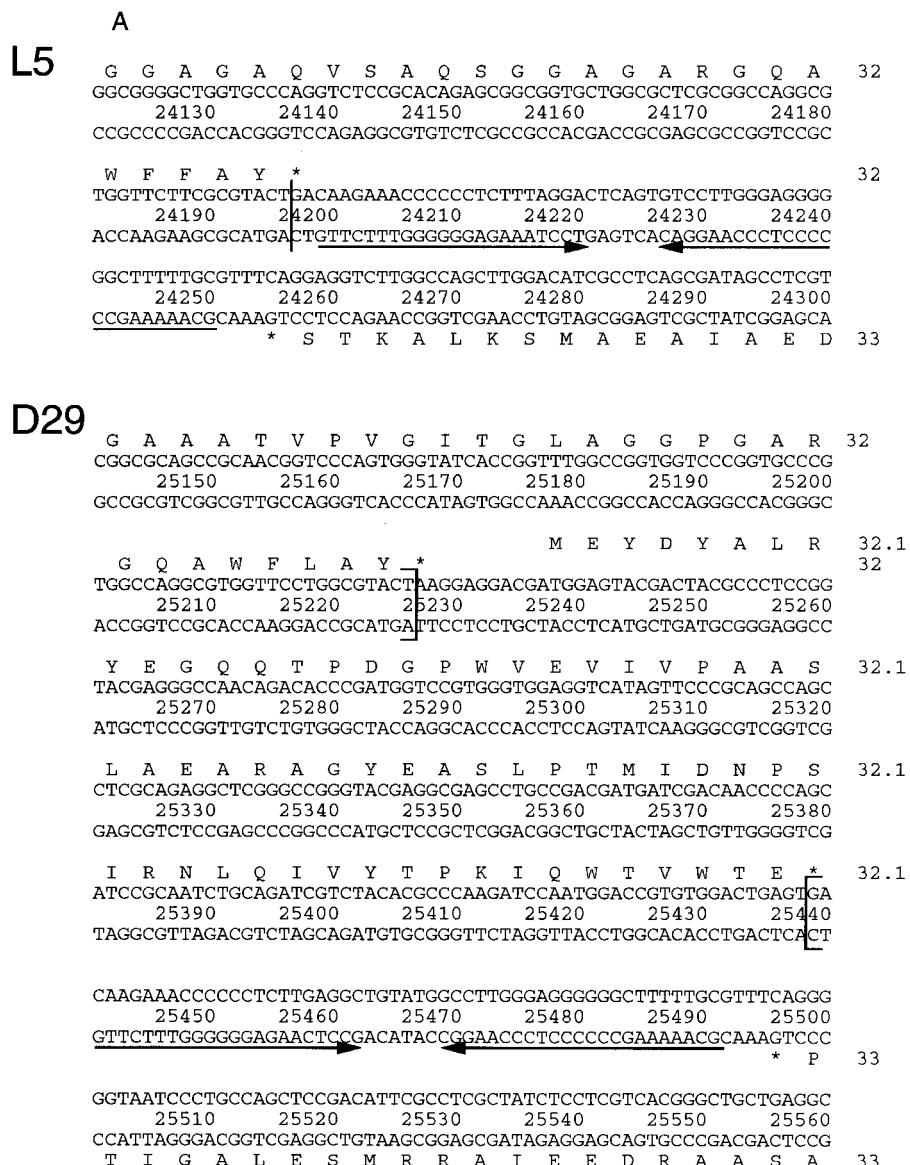


Figure 7A (legend opposite)

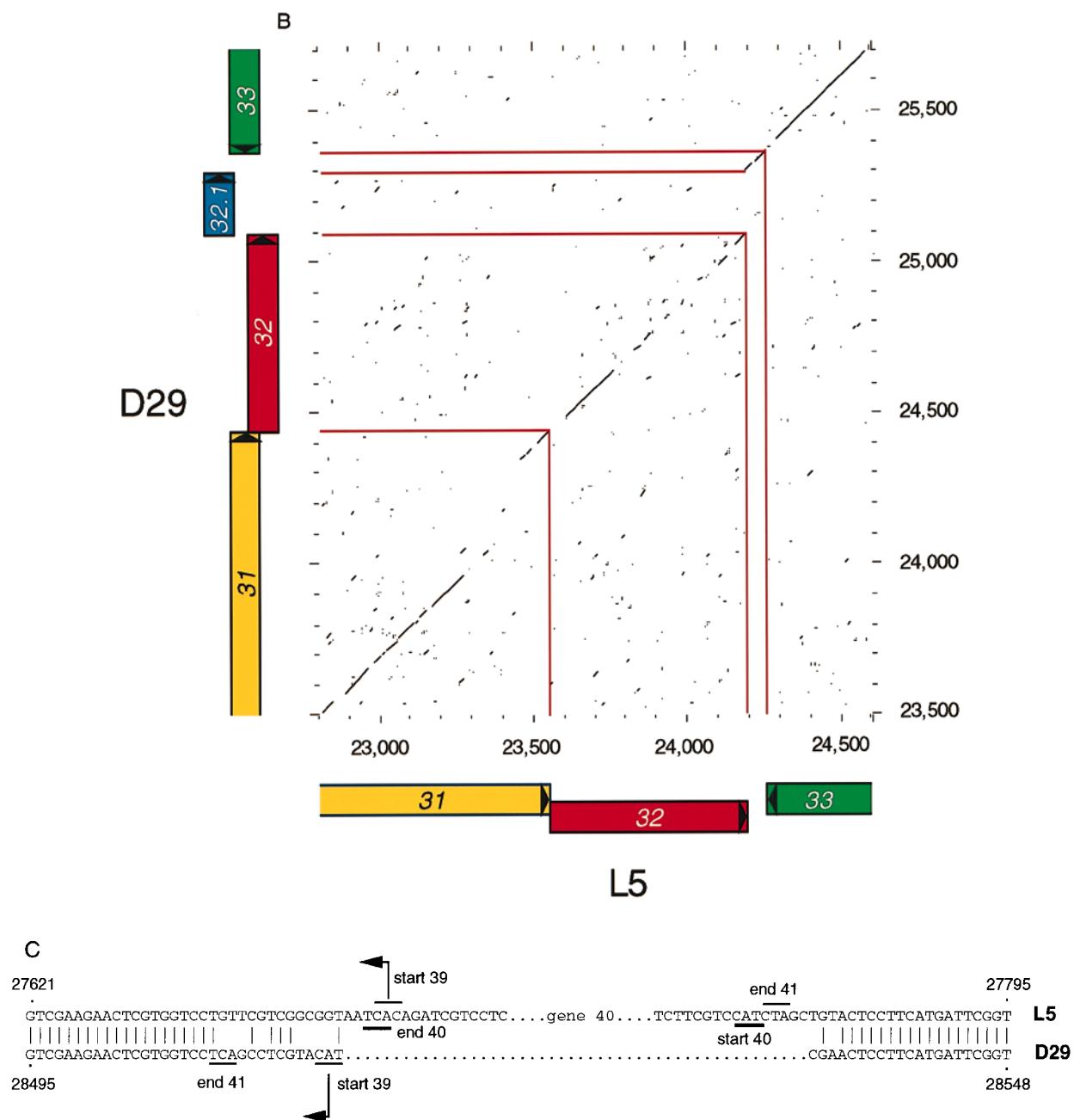


Figure 7. Insertions and deletions in L5 and D29. A, Comparison of the ends of the late lytic operons of L5 and D29 showing the presence in D29 of DNA not possessed by L5. Immediately following L5 gene 32 and separating it from gene 33 on the opposite strand is an imperfect 22 bp inverted repeat which could potentially fold into a stem-loop terminator-like structure. A very similar inverted repeat structure is found in D29, but in this case it follows not gene 32, but gene 32.1, which is not found in L5. The deletion of the 212 bp outside of the brackets of a D29-like precursor could give rise to the L5 arrangement. B, Diagon plot comparing the regions of the L5 and D29 chromosomes described in A. Note that gene 32.1 is absent from L5. However, the intergenic space containing the stem-loop terminator-like structure is conserved for both phages. C, Sequencing alignment of the gene 39 through 41 regions of L5 and D29 showing the presence of L5 of a gene not found in D29. Initiation and termination codons are indicated. In this case, L5 contains a gene that D29 does not: gene 40. A relatively simple deletion of DNA from an L5-like ancestor could produce the D29 arrangement.

41 region that is illustrated in Figure 8A. Whereas both L5 and D29 contain homologous copies of genes 36 (89.3% amino acid identity) and 41 (81.9% amino acid identity) and related copies of gene 39 (54.9% amino acid identity), the DNA sequences between genes 36 and 39 appear unrelated. While amino acid comparisons suggest that the L5 and

D29 products of gene 38 are related (36.7% amino acid identity) this is not the case for L5 gene 37 and D29 gene 36.1.

Previous analysis of L5 gene 37 failed to identify any relatives by database searching and a function could not be assigned. In contrast, D29 gene 36.1 (which is very similar in size to L5 37) apparently

encodes a deoxycytidylate deaminase (dCMPase) with reasonable sequence similarity (average amino acid identity, 19%) to the human, yeast, *Bacillus subtilis*, and bacteriophage T2 and T4 enzymes. The T4 dCMPase is involved in nucleotide metabolism and converts dCMP to dUMP which is subsequently converted by another enzyme into dTTP (Greenberg *et al.*, 1994). The D29 enzyme may fulfill a similar function. We do not yet know if D29 36.1 (or L5 37) is essential for viral growth but note that there are several other genes in the right arm that encode proteins involved in nucleotide metabolism or DNA synthesis (see Figure 3).

A second example of gene substitution is D29 gene 66.1, which substitutes for L5 gene 67. The putative recombination event responsible for this gene replacement occurred close to, but not directly at, the gene boundaries. At the DNA sequence level of comparison, the two phage genome sequences diverge approximately 20 bp prior to the end of gene 68, leading to predicted proteins with dissimilar C termini. Likewise, strong DNA sequence similarity resumes approximately 30 bp prior to the end of the non-matching genes, but different reading frames prevent identical translations. The translations begin matching again at the beginning of genes 66 (Figure 8B). Therefore, the predicted gene products of L5 67 and D29 66.1 are very similar in size but essentially unrelated in sequence. The function of neither gene product is known. Because the recombination that is presumed to have given rise to this difference between D29 and L5 did not occur precisely at the endpoints of the affected gene, it provides evidence for the assertion that the illegitimate recombination events that are presumed to lead to mosaic-like arrangements of phage genomes can occur anywhere within the genome, but only survive to be analyzed if they do not disrupt any essential functions.

A third example is the substitution of D29 genes 59.1 and 59.2 for L5 gene 60 (Figure 8C). The size of the substituting DNA is quite different, with a segment of 923 bp in D29 replacing 231 bp in L5.

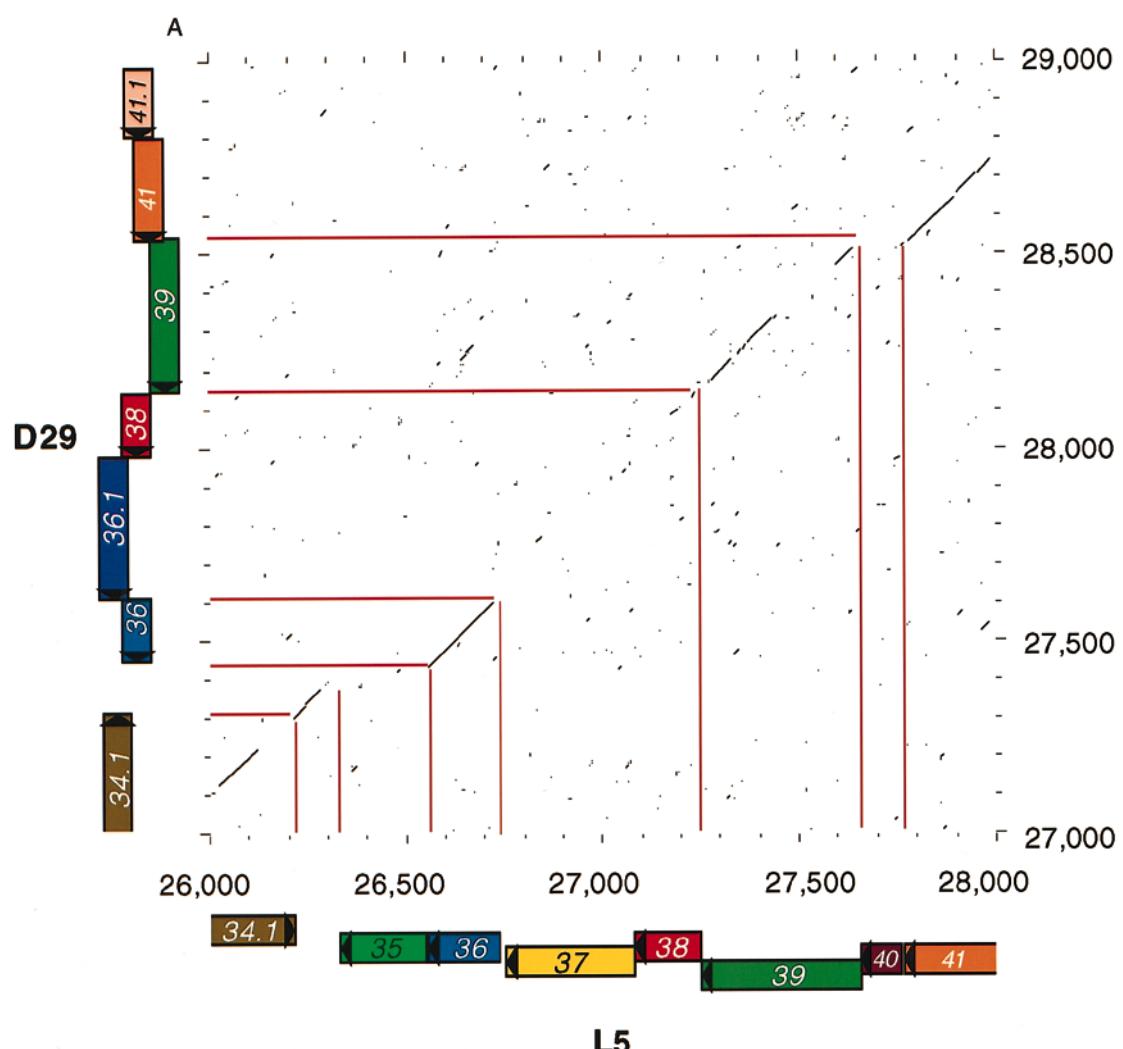


Figure 8A (legend on page 158).

However, the boundaries are quite well-defined, positioned precisely at the initiation codons for gene 59 and the termination codons for gene 61 (coordinates 39,287 and 39,518 in L5; 39,621 and 40,549 in D29; the intervening segment in L5 contains gene 60, and in D29 contains genes 59.1 and 59.2). While the functions of L5 gene 60 and D29 59.2 are not known, D29 59.2 appears to encode a member of the family of bacterial non-heme haloperoxidases, with about 25% amino acid identity to its closest relative (the CPO-C enzyme of *Pseudomonas pyrrocinia*; Wolfframm *et al.*, 1993) and conservation of the catalytic triad that has been established for this family of enzymes (Pelletier *et al.*, 1995). The product of gene 59.2 is the only

virally encoded member of this protein family of which we are aware.

The two remaining substitutions are similar in nature to those discussed above. D29 gene 44.1 substitutes for L5 gene 45 with the sequence departures at the initiation and termination codons of the flanking genes (D29 coordinates 31,162 to 31,506 replace L5 coordinates 30,476 to 30,962). A particular consequence of this substitution is that translation initiation of D29 gene 44 occurs at an AUG codon, whereas L5 gene 44 starts with a UUG codon. This further supports the idea that UUG is used for translation initiation in the mycobacteria. Similarly, genes, 82.1 and 82.2 substitute for L5 gene 83 with

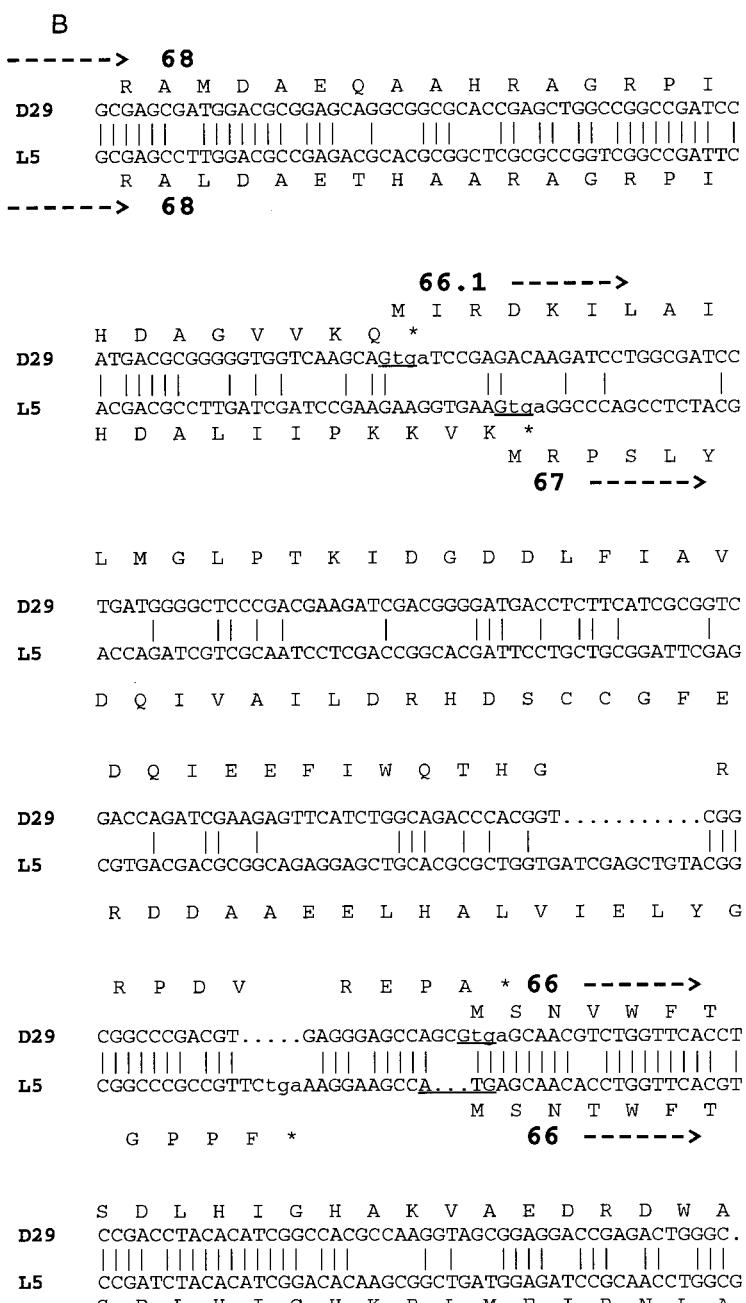


Figure 8B (legend on page 158).

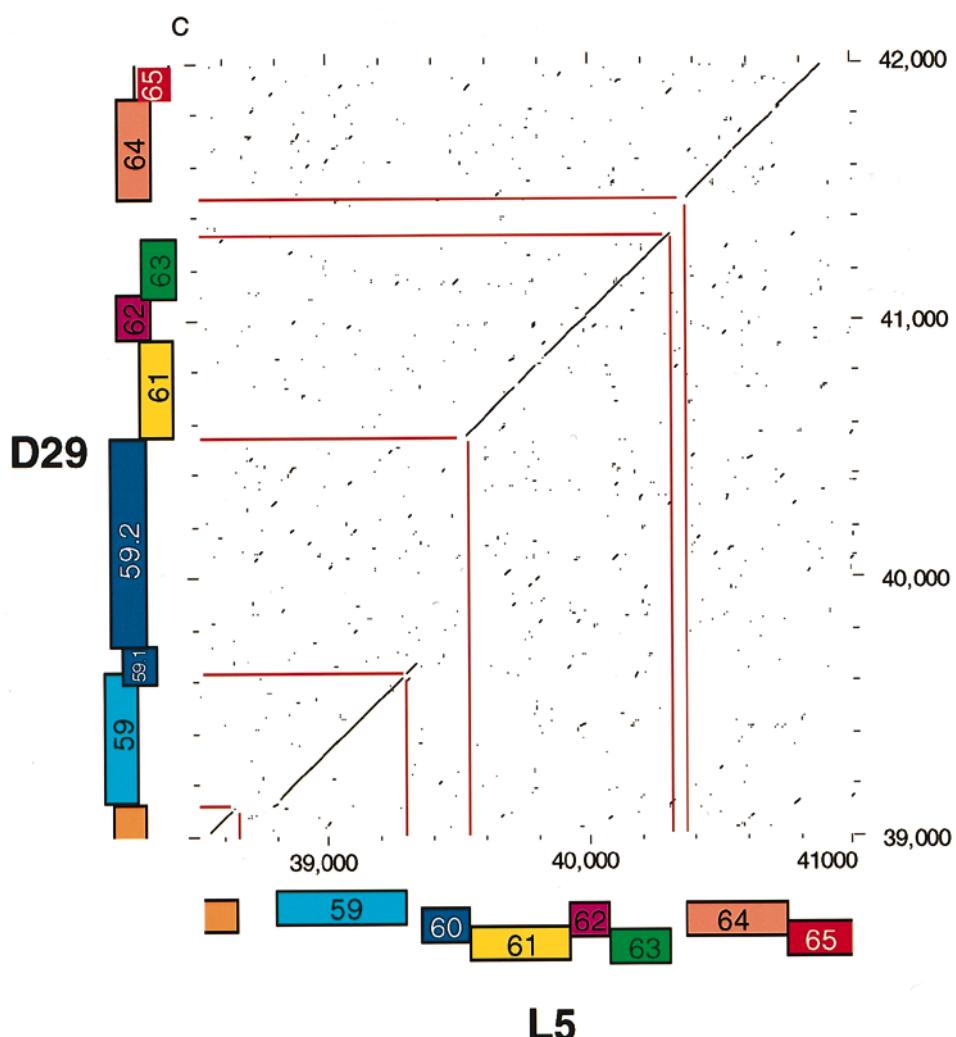


Figure 8. Gene substitutions in L5 and D29. A, Diagon plot comparison of the gene 36 to 41 regions of L5 and D29. Homologous genes 36 and related genes 39 flank a stretch of DNA that is quite different between the two phages. L5 gene 37 has been substituted in D29 by gene 36.1, the product of which is potentially a dCMP deaminase, and genes 38 share only limited similarity. Note also that the D29 gene 31 contains additional DNA near the 3' end which is absent from the L5 homolog. B, DNA alignments and amino acid translations of D29 and L5 across a region of their chromosomes which represents a gene replacement. Portions of the two flanking genes are shown. The C-terminal ends of common genes 68, which share a 68.4% amino acid identity along their lengths, and the N-terminal ends of genes 66 (65.2% amino acid identity) are shown flanking gene 66.1 in D29 and gene 67 in L5. Although these two genes are situated identically and are of nearly the same length, discernible sequence similarity is absent. Initiation codons are underlined and termination codons are in lower case. Dots indicate spaces that have been introduced into one of the DNA sequences in order of maximize alignment. C, Diagon plot comparison of the L5 and D29 genome regions surrounding another apparent gene replacement. Note that the relatively small L5 gene 60 has been replaced in D29 by a larger piece of DNA containing two genes, 59.1 and 59.2. The functions of L5 and gene 60 and D29 gene 59.1 are unknown, but D29 gene 59.2 potentially encodes a non-heme haloperoxidase.

sequence departures close to the initiation and termination codons of the flanking genes (D29 coordinates 45,905 to 46,118 replacing L5 coordinates 48,519 to 48,728).

All of the substitutions discussed above are readily observed since the substituting sequences are quite different at both the DNA and protein level. However, in addition to D29 gene 38 described previously, there are several other D29 genes (30, 49, 55, 70 and 84) that show a degree of similarity to their L5 counterparts but are more distantly related than most L5/D29 pairs of gene

products. Each of these six gene products is less than 50% identical to its L5 homolog, whereas the average identity between L5 and D29 gene products is 74%.

Signals for gene expression and regulation

The sequence requirements for mycobacterial promoters are not well-defined (Bashyam *et al.*, 1996) and as yet it is not possible to positively identify them from sequence analysis alone. The only known L5 promoters are P1, P2, and P3

located immediately upstream of gene 71, and the early lytic promoter, P_{left} , which is located at the extreme right end of the L5 genome and transcribes leftwards (Figure 3). The gene 71 promoters lie within the large region of L5 DNA that has been deleted from D29 and are therefore not present in this phage. Comparison of the D29 and L5 sequences at P_{left} shows that they are very similar (Figure 9A). However, the putative -10 region appears to be poorly conserved. We do not know if this difference influences the activity of this promoter. The P_{left} promoter of L5 is regulated by the gp71 repressor that binds to a 13 bp operator site overlapping the -35 region (Brown *et al.*, 1997). This site is also present in D29, with one difference

from the L5 site that probably has little or no effect on the binding affinity of gp71 (Figure 9; Brown *et al.*, 1997).

A novel feature of the L5 genome is the presence of multiple asymmetric gp71-binding sites located throughout the genome. These are predominantly within short intergenic or non-coding regions and are oriented in one direction relative to transcription. They have been proposed to act as "stopoperator" sites, whereby the binding of gp71 results in transcription termination and down-regulation of gene expression in the prophage. We have therefore searched for similar sites within the D29 genome. We find that most of the sites in L5 are also present in D29 (in similar positions and ori-

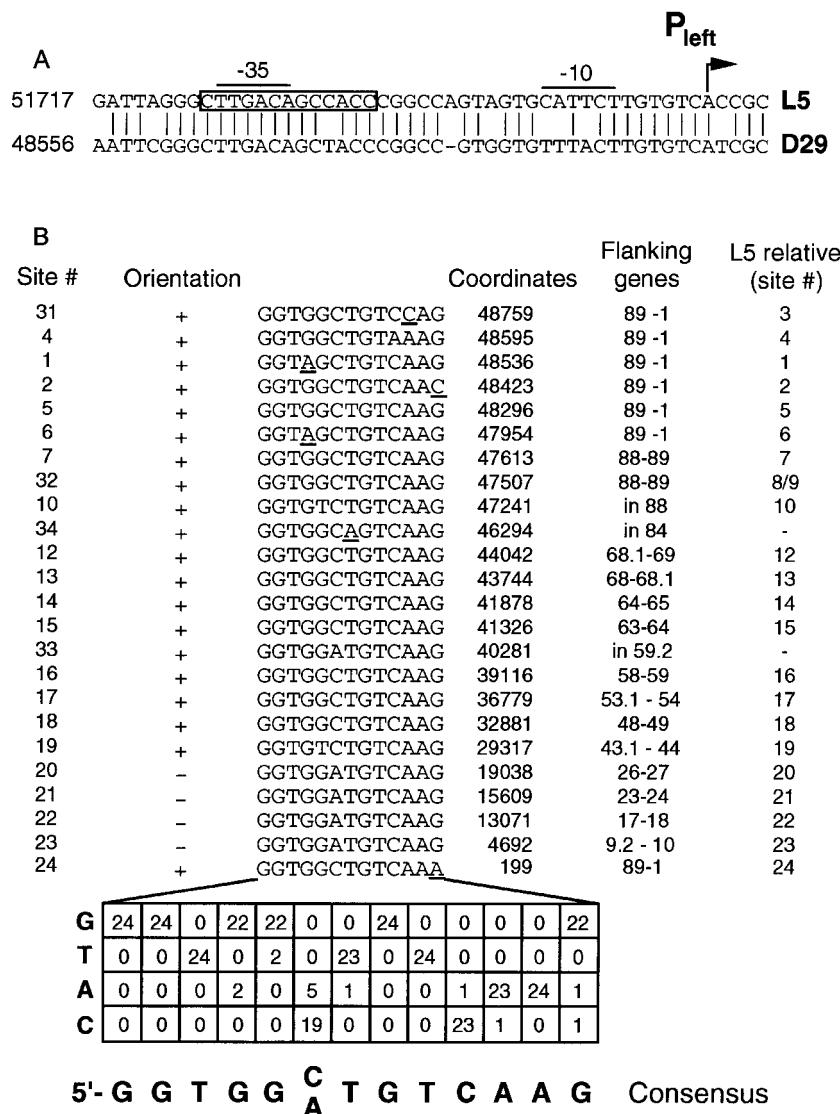


Figure 9. Putative repressor binding sites in D29. A, The DNA sequences of L5 and D29 are aligned at the position of the early lytic promoter P_{left} . The position of transcription initiation in L5 is shown by the arrow and the putative -10 and -35 regions are indicated. The repressor binding site upstream of P_{left} is boxed. Note that the P_{left} promoter is located at the right end of the phage genome and transcribes leftwards (see Figure 3). B, 24 putative repressor binding sites of D29 are aligned and a consensus sequence presented. The coordinates of the sites, orientations in the genome, flanking genes and corresponding sites in L5 are shown. See Figure 3 for further information on location and orientation of the sites.

tations; Figure 3), and conform to the same consensus as that for the L5 sites (5'-GGTGGc/aTGT-CAAG; Figure 9B).

Phage L5 has several apparent factor-independent (stem-loop) type terminators, all located close to *attP* at the center of the genome (see Figure 3). The terminator to the left of *attP* is located at the ends of genes 32 and 33 where the transcriptional units converge (see Figure 7A); it could function as a terminator in both directions and is a candidate for retro-regulation of *int* expression (Lee *et al.*, 1991). The sequence is very similar in D29 and can form similar putative secondary structures. L5 has a putative rightwards terminator that is positioned immediately to the right of the common core in *attP*, although this is not conserved in D29 (Peña *et al.*, 1997). The sequence composing the putative leftwards terminator at the end of L5 gene 35 (a 17 bp stem followed by seven U-residues) is not present in D29, but the sequence at the end of D29 gene 36 can fold into a similar terminator-like structure. It seems probable that at least the conserved terminators have important regulatory functions.

Other sequence features

There are several other noteworthy features of the D29 genome. First, although none of the deletions/insertions discussed can be simply accounted for by recombination between short stretches of similar sequences, there is at least one instance where this appears to have occurred. In D29, the region between gene 88 and 89 is about 270 bp shorter than it is in L5 and contains only two gp71-binding sites as opposed to the three in L5. Deletion of the region between binding sites 8 and 9 could have given rise to the organization observed in D29.

It was previously suggested that L5 genes 24 and 25 are organized such that they could be expressed *via* a programmed translational frame-shift (Hatfull & Jacobs, 1994; R. W. Hendrix, personal communication). This mechanism was first demonstrated for the lambda genes G and T (Levin *et al.*, 1993), which occupy similar positions as 24 and 25 relative to other structural genes. In D29, genes 24 and 25 are organized similarly to L5 and may also be expressed *via* a translational frame-shift. The amino acid sequence of the region between the proposed site of the frameshift (eight to nine codons from the 3' end of 24), and the first plausible translation initiation site for D29 gene 25 (approximately 140 bp downstream) are very similar to the L5 sequence (85% identity), an observation that provides further support for the frameshifting model.

Discussion

We have presented here the sequence analysis and genome map of mycobacteriophage D29. This is only the second complete mycobacteriophage

genome sequence to be reported and is clearly a close relative of the previously sequenced L5 genome (Hatfull & Sarkis, 1993). D29 thus serves as a useful comparison for the L5 genome map and provides considerable information about the biology of this group of bacteriophages.

An obvious difference between the L5 and D29 genomes is the large deletion in the right arm of D29 that removes part of the repressor gene. This deletion fully accounts for the lytic phenotype of the D29 isolate that we sequenced. However, a number of features of the D29 genome suggest that this event occurred relatively recently, perhaps shortly after the isolation of D29. For example, D29 possesses all of the required functions for integration including the *attP* site and integrase gene (33), both of which appear to be functional (J. Stoner & G.F.H., unpublished observations). In addition, many of the repressor binding sites that are found in L5 are also present in D29, at least some of which can be bound by L5 gp71 *in vitro* (C. Wadsworth, K. Brown & G.F.H., unpublished observations). Finally, we have demonstrated that D29 is capable of lysogenizing *M. smegmatis* when L5 gp71 is provided from an extrachromosomal plasmid. Besides the current genomic evidence, there are past reports that seem to support the conclusion that D29 is recently derived from a temperate parental phage. Work on the original isolate of D29 noted the appearance of a mixed plaque morphology when the phage was plated on a mycobacterial lawn. Some of these plaques were turbid (Bowman, 1958). In addition, although the identity of the temperate phage responsible was unclear, two groups from this same time period reported the isolation of D29 lysogens of *M. smegmatis* (Russell *et al.*, 1963; Tokunaga & Sellers, 1970).

Phages D29 and L5 are somewhat unusual (among temperate phages) in carrying genes involved in nucleotide metabolism. Prior to our sequence analysis of the D29 genome, we were aware only of gene 50, which encodes a B12-dependent ribonucleotide reductase of which the only close relative is an enzyme encoded by *Lactobacillus leichmannii* (Booker & Stubbe, 1993), although gp56 encodes a glutaredoxin-like protein that could act in conjunction with gp50. The finding that D29 also encodes a dCMP deaminase suggests that wholesale modification of the intracellular nucleotide pools occurs during lytic growth. Both phages also encode other DNA replication functions (including a DNA polymerase), so these features may reflect a need to rapidly synthesize DNA under conditions where cellular DNA replication is either not occurring, or the host enzymes in both nucleotide metabolism and DNA synthesis are severely limiting. Little is known about the processes of DNA replication in the mycobacteria, and many members of the genus grow extremely slowly, making this adaptation on the part of L5 and D29 particularly interesting. The coliphage T4 also encodes a dCMP

deaminase which plays a specific role in the synthesis of dTTP (Greenberg *et al.*, 1994). It has been argued that the synthesis of dTTP is important since the A + T content of T4 (~65%) is substantially higher than that of *E. coli*, and that a host-encoded dCMP deaminase serves to ensure an adequate supply of dTTP (Greenberg *et al.*, 1994). However, this scenario is not applicable to D29 since the G + C content is relatively high (63%), similar to that of its hosts.

The role of the tRNA genes in phages L5 and D29 is not known. However, two of the D29 tRNA genes (9.1 and 9.2) may not be required for growth since they are absent from L5. We have shown that at least two of the L5 tRNA genes (8 and 9) are functional since the tRNAs they encode can be mutated to active nonsense suppressors (M. Tang, G.F.H. & C. Peebles, unpublished observations) and all of the D29-encoded tRNAs have reasonably normal tRNA structures (Figure 5B). None of these tRNAs recognize rare codons and they may simply boost the level of protein synthesis during lytic growth. The fact that D29 contains five tRNA-like genes (and L5 three) indicates the existence of an ancestral mycobacteriophage "tRNA cassette" that contained at least the five tRNA genes found in D29 and possibly others. We would predict that other members of the L5-like family encode tRNAs as well. It would be interesting to know which tRNA genes are encoded by different members of this group, particularly those of markedly different host range than L5 or D29. In addition, it would be interesting to know if other mycobacteriophages that are not related to L5 encode their own tRNAs, in order to ascertain if infection of the mycobacteria by phage requires that the phage provide its own tRNAs, or if this is a feature peculiar to the growth of the L5-like family.

Both the L5 and D29 genomes contain a multitude of repressor binding sites, situated in direct orientation at gene boundaries. Although these are small (13 bp) it is not unreasonable to suppose that they could act as hot spots for recombination to produce gene insertions, deletions or substitutions. However, while many such differences are observed between L5 and D29, none appear to be generated *via* this mechanism. The only possible example of such a recombination event is at the right end of the D29 genome where a 268 bp segment between L5 gp71 binding sites 8 and 9 is absent. Thus the plethora of observed gene substitutions and deletions seen between L5 and D29 must have occurred by other mechanisms involving illegitimate recombination. The slow-growing mycobacteria such as *M. tuberculosis* have been shown to have a rather high frequency of illegitimate recombination (and only inefficient homologous recombination) which could promote these events (McFadden, 1996).

There are several instances where L5 genes have apparently been replaced in D29 by a novel gene or genes. Examples of gene replacements

between two otherwise related phages are most easily explained if they are the result of a recombination at a gene boundary between two related phages, and this is the explanation that we favor. Recombination between phages has been postulated to be a more effective mechanism of phage evolution if the recombining genomes share the same overall genetic organization (Casjens *et al.*, 1992). Therefore, individual gene position appears to be important and it is often the case amongst lambdoid phages that two distinct genes whose products accomplish the same function by different catalytic mechanisms are situated identically within the phages' genomes. One well known example is the R lysis genes of lambda and P22. The P22 enzyme is a true lysozyme (Rennell & Poteete, 1985), while lambda encodes a transglycosylate (Bienkowska-Szewczyk & Taylor, 1980). Both enzymes serve to cleave the polysaccharide portion of the bacterial murein layer, but they accomplish this task by different catalytic mechanisms. It is thus tempting to speculate that this may be the case for the L5-like phages as well. Interestingly, the products of the L5 and D29 gene 30, which are only 38.8% identical to each other, are potentially involved in lysis of the mycobacterial cell (K. J. Fullner & G.F.H., unpublished observations). Perhaps other genes of these two phages that are quite different from one another yet situated identically within the genomes encode analogous functions.

In addition to the information provided regarding mycobacteriophage biology, the complete DNA sequence of D29 has enabled an evolutionary comparison of two closely related mycobacteriophages, and this comparison has allowed us to extend some of the lambdoid bacteriophage evolution principles to a superficially unrelated group of viruses, the mycobacteriophages. The genomes of L5 and D29, while very closely related at the nucleotide level of comparison, are punctuated by a large number of discontinuities such as insertions, deletions, and gene replacements. Thus, the genomes of L5 and D29 appear to be genetic mosaics much like those of lambdoid phages. While the degree of mosaicism in L5 and D29 is less than that seen with the lambdoid phages, this probably reflects their evolutionary proximity rather than a fundamental difference among phage groups. Comparison with additional mycobacteriophage genomes will be necessary to address these issues.

Methods and Materials

Bacteria and phages

Phage D29 was provided by Dr W. R. Jacobs Jr, Albert Einstein College of Medicine, New York, who obtained it from Dr W. Jones. *M. smegmatis* strain mc²155 was from a laboratory stock.

Phage purification and DNA isolation

D29 phage particles were isolated using a standard plate lysate procedure (Sambrook *et al.*, 1989). Approximately 2×10^4 PFUs were mixed with 1 ml of late-log phase *M. smegmatis* mc²155 cells and 9 ml of 7H9 top agar (Middlebrook 7H9 broth base, Difco Laboratories, Detroit, MI + 0.75% agar) supplemented with 1 mM CaCl₂. This mixture was plated on a 150 mm Petri dish containing Middlebrook 7H10 agar (Difco Laboratories, Detroit, MI) supplemented with 1 mM CaCl₂. (For all phage manipulations, oleic acid and Tween 80 were omitted from the bacterial growth medium). Ten such plates were prepared and incubated overnight at 37°C, after which time a slight bacterial "webbing" was seen indicating nearly confluent bacterial lysis by the phage. Phage particles were collected by addition of 10 ml of phage buffer (10 mM Tris-HCl (pH 7.5), 10 mM MgSO₄, 68.5 mM NaCl) to the surface of each plate. After a four hour incubation at 4°C, the phage-containing buffer was pipetted off the plate. Bacterial cells were removed by centrifugation (9000 g for ten minutes), and the phage remaining in the supernatant were precipitated by the addition of polyethylene glycol (PEG) to 10% and NaCl to 1 M, followed by centrifugation as above. The phage-containing pellet was resuspended in a small volume of phage buffer and the phage virions were purified using CsCl density gradient centrifugation (Sambrook *et al.*, 1989). CsCl was removed by dialyzing repeatedly against phage buffer. DNA was isolated from the purified phage particles by sequential phenol/chloroform/isoamyl alcohol (25:24:1, by vol) extractions until the interface was clean. The DNA was then ethanol precipitated and resuspended in TE buffer.

DNA preparation and sequencing

The D29 chromosomal DNA library was prepared using an adaptation of a previously published protocol (Démolis *et al.*, 1995). The purified phage DNA was partially digested with an appropriate amount of DNase I (Boehringer Mannheim, Indianapolis, IN; usually 10 to 15 µg of DNA and 0.015 to 0.030 units DNase I/µg of DNA) in a buffer consisting of 50 mM Tris (pH 7.0), 10 mM MnCl₂ for one minute at room temperature. The DNA was repaired with Klenow and T4 DNA polymerases (New England Biolabs, Inc., Beverly, MA) plus 500 µM dNTPs for 30 minutes at room temperature and size-fractionated on a 0.7% (w/v) agarose gel. The 1 to 3 kb size fraction was isolated from the gel and the corresponding fragments cloned into the EcoRV site of pBluescript SK-(Stratagene, La Jolla, CA). The resulting ligation mixture was used to transform *E. coli* XLI-Blue (Stratagene, La Jolla, CA) by electroporation (Sambrook *et al.*, 1989), and insert-containing white colonies were grown overnight in LB medium in preparation for plasmid isolation. The double-stranded DNA clones were purified using the Qiagen QIAwell 96 Ultra exchange DNA purification kit (Qiagen, Inc., Santa Clarita, CA). They were sequenced from both ends on the Perkin Elmer ABI Prism 377 DNA Sequencer using the Dye Terminator chemistry (Perkin Elmer, Foster City, CA). The sequence was completed using a small number of primers constructed to fill in gaps and either a double-stranded clone or whole D29 genome as template DNA. Long Ranger 5% gels (FMC BioProducts, Rockland, ME) were used in place of the standard 4% polyacrylamide gel recommended by Perkin Elmer.

Sequence assembly and analysis

Random phage sequences obtained from the automated sequencer were assembled and edited using Sequencher software (Gene Codes, Ann Arbor, MI). Sequence analysis was performed using Staden (1986), GeneMark (Borodovsky & McIninch, 1993), and GCG (Genetics Computer Group, Madison, WI) programs. The use of a previously constructed L5 codon usage table facilitated the identification of D29 open reading frames (Hatfull & Sarkis, 1993). The complete DNA sequence of mycobacteriophage D29 can be found in GenBank under accession number AF022214.

Isolation and characterization of mycobacteriophage D29 lysogens

Aliquots (10 µl) of 100-fold serial dilutions of phage stocks (1×10^{10} PFUs/ml) of the D29::FFlux reporter phage phBD8 (Pearson *et al.*, 1996), and the L5::FFlux reporter phages pGS18 and pGS26 (Sarkis *et al.*, 1995) were spotted onto a lawn of *M. smegmatis* mc²155 containing the plasmid pMD132 (Donnelly-Wu *et al.*, 1993; pMD132 is a shuttle plasmid with a 1.3 kb fragment of L5 DNA containing parts of genes 70 and 72 and all of gene 71). Cells from the center of spots (containing approximately 1×10^8 phage) for each phage were streaked onto solid medium (Middlebrook 7H10 agar supplemented with 1 mM CaCl₂) to obtain single colonies. Approximately 50 isolated colonies from each infection were transferred onto a fresh plate overlaid with a nitrocellulose membrane (Protran[®]; Schleicher & Schuell, Keene, NH) and incubated for three days at 37°C. Colonies were photographed directly under incident light. Photon emission from the colonies was photographed after the addition of 1 ml of 1 mM d-luciferin, 100 mM sodium citrate (pH 5.0).

Cells from glowing colonies (five of each strain) were streaked onto fresh medium and isolated colonies were grown in 1.5 ml of Middlebrook 7H9 broth to an *A*₆₀₀ of approximately 1.0. Cells were pelleted by centrifugation from 1 ml of culture, resuspended in 100 µl of TE, and boiled for five minutes. PCR was performed essentially as described by Peña *et al.* (1997) and using the same primers. The primers OL-316 (5'-GCTGCCATGCGAAA-CAGGCT) and OL-317 (5'-AAAACCACCTCTGACCT-GTG), which flank the D29 *attP* site, were designed to replace the L5 *attP* primers in PCR reactions involving phBD8 such that the PCR products from phBD8 (*attP*) and lysogens of phBD8 (*attL* and *attR*) would be similar to those of L5 and L5 lysogens. The D29 *attP* primers do not amplify L5 *attP* and the L5 *attP* primers do not amplify D29 *attP*.

SDS-PAGE analysis of phage particles

Approximately 50 µl of a 10^{12} PFUs/ml cesium chloride-purified stock of D29 or L5 were centrifuged to pellet the phage. The phage pellet was then resuspended in 75 µl of water, vortexed to mix, and frozen at -70°C. The frozen mixture was rapidly thawed and mixed by vortexing. This process was repeated twice, and the mixture was then heated to 75°C for three to four minutes. After heating, the phage suspension had become quite viscous, indicating that the chromosomal DNA had been released from the phage virions. DNaseI (20 units, Boehringer Mannheim, Indianapolis, IN) was added and the solution was allowed to incubate at 37°C for 30 to 60 minutes, by which time the viscosity had diminished.

Approximately 25 µl of 4 × SDS protein sample buffer (Sambrook *et al.*, 1989) were added and the solution was boiled for 2.5 minutes: 15 µl of the resulting solution was electrophoresed through an SDS 10% polyacrylamide gel (polyacrylamide:bisacrylamide, 30%:80%). Proteins were visualised by staining the gel with Coomassie brilliant blue dye.

Determination of N-terminal amino acid sequences

The D29 proteins from a gel similar to the one shown in Figure 2 were transferred to PVDF paper (BioRad, Hercules, CA) and stained with Coomassie brilliant blue dye. Paper strips containing isolated protein bands were excised (LeGendre & Matsudaira, 1989), and N-terminal sequence analysis was performed in a protein sequenator (Porton 2090E, Beckman Instruments, Inc., Fullerton, CA).

Acknowledgements

We thank David Stone for technical assistance, Greg Morgan for critical comments on the manuscript, John Hempel for protein sequence analysis, Tom Harper for help with electron microscopy, Robert Duda for assistance with phage protein preparation, and Robert Suto for bringing to our attention the identification of L5 gp50 as a ribonucleotide reductase. This work was supported by NIH grant GM51975.

References

- Bardarov, S., Kriakov, J., Carriere, C., Shengwei, Y., Vaamonde, C., McAdam, R. A., Bloom, B. R., Hatfull, G. F. & Jacobs, W. R. (1997). Conditionally replicating mycobacteriophages: A system for transposon delivery to *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA*, **94**, 10,961–10,966.
- Barsom, E. K. & Hatfull, G. F. (1996). Characterization of a *Mycobacterium smegmatis* gene that confers resistance to phages L5 and D29 when overexpressed. *Mol. Microbiol.* **21**, 159–170.
- Bashyam, M. D., Kaushal, D., Dasgupta, S. K. & Tyagi, A. K. (1996). A study of mycobacterial transcriptional apparatus: identification of novel sequence features in promoter elements. *J. Bacteriol.* **178**, 4847–4853.
- Besra, G., Khoo, K.-H., Belisle, J. T., McNeil, M. R., Morris, H. R., Dell, A. & Brennan, P. J. (1994). New pyruvylated, glycosylated acyltrehaloses from *Mycobacterium smegmatis* strains, and their implication for phage resistance in mycobacteria. *Carbohydr. Res.* **251**, 99–114.
- Bienkowska-Szewczyk, K. & Taylor, A. (1980). Murein transglycosylate from phage lambda lysate. Purification and properties. *Biochim. Biophys. Acta*, **615**, 489–496.
- Bloom, B. R. & Murray, C. J. L. (1992). Tuberculosis: commentary on a reemergent killer. *Science*, **257**, 1055–1064.
- Booker, S. & Stubbe, J. (1993). Cloning, sequencing, expression of the adenosylcobalamin-dependent ribonucleotide reductase from *Lactobacillus leichmanii*. *Proc. Natl Acad. Sci. USA*, **90**, 8352–8356.
- Borodovsky, M. & McIninch, J. D. (1993). GeneMark: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17**, 123–133.
- Bowman, B. U., Jr (1958). Quantitative studies on some mycobacterial phage-host systems. *J. Bacteriol.* **76**, 52–62.
- Broida, J. & Abelson, J. (1985). Sequence organization and control of transcription in the bacteriophage T4 tRNA region. *J. Mol. Biol.* **185**, 545–563.
- Brown, K. L., Sarkis, G. J., Wadsworth, C. & Hatfull, G. F. (1997). Transcriptional silencing by the mycobacteriophage L5 repressor. *EMBO J.* **16**, 5914–5921.
- Casjens, S., Hatfull, G. & Hendrix, R. (1992). Evolution of dsDNA tailed-bacteriophage genomes. *Seminars in Virology* (Koonin, E., ed.), vol. 3, pp. 383–397, Academic Press, London.
- Clark-Curtiss, J. (1990). Genome structure of the mycobacteria. In *Molecular Biology of the Mycobacteria* (McFadden, J., ed.), pp. 77–96, Academic Press, London.
- David, H., Clement, F., Clavel-Seres, S. & Rastogi, N. (1984). Abortive infection of *Mycobacterium leprae* by the mycobacteriophage D29. *Int. J. Lep.* **52**, 515–523.
- Démolis, N., Mallet, L., Bussereau, F. & Jacquet, M. (1995). Improved strategy for large-scale DNA sequencing using DNaseI cleavage for generating random subclones. *BioTechniques*, **18**, 453–457.
- Doke, S. (1960). Studies on mycobacteriophages and lysogenic mycobacteria. *J. Kumamoto Med. Soc.* **34**, 1360–1373.
- Donnelly-Wu, M. K., Jacobs, W. R., Jr & Hatfull, G. F. (1993). Superinfection immunity of mycobacteriophage L5: Applications for genetic transformation of mycobacteria. *Mol. Microbiol.* **7**, 407–417.
- Duda, R. L., Hempel, J., Michel, H., Shabanowitz, J., Hunt, D. & Hendrix, R. W. (1995). Structural transitions during bacteriophage HK97 head assembly. *J. Mol. Biol.* **247**, 618–635.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocajne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Sudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
- Froman, S., Will, D. W. & Bogen, E. (1954). Bacteriophage active against virulent *Mycobacterium tuberculosis* I. Isolation and activity. *Am. J. Pub. Health*, **44**, 1326–1333.
- Fullner, K. J. & Hatfull, G. F. (1997). Mycobacteriophage L5 infection of *Mycobacterium bovis* BCG: implications for phage genetics in the slow-growing mycobacteria. *Mol. Microbiol.* **26**, 755–766.
- Greenberg, R. G., He, P., Hilfinger, J. & Tseng, M. (1994). Deoxyribonucleoside triphosphate synthesis and phage T4 DNA replication. In *Molecular Biology of Bacteriophage T4* (Karam, J. D., ed.), pp. 14–27, ASM Press, Washington, DC.

- Hatfull, G. F. (1994). Mycobacteriophage L5: A toolbox for tuberculosis. *ASM News*, **60**, 255–260.
- Hatfull, G. F. & Jacobs, J. R., Jr (1994). Mycobacteriophages: cornerstones of mycobacterial research. In *Tuberculosis: Pathogenesis, Protection and Control* (Bloom, B. R., ed.), pp. 165–183, ASM, Washington DC.
- Hatfull, G. F. & Sarkis, G. J. (1993). DNA sequence, structure, and gene expression of mycobacteriophage L5: A phage system for mycobacterial genetics. *Mol. Microbiol.* **7**, 395–405.
- Hatfull, G. F. (1994). Genetics of *Mycobacterium tuberculosis*. Section III. In *Tuberculosis: Pathogenesis, Protection, and Control* (Bloom, B. R., ed.), pp. 165–268, ASM Press, Washington, DC.
- Jacobs, W. R., Jr, Barletta, R. G., Udani, R., Chan, J., Kalkut, G., Sosne, G., Kieser, T., Sarkis, G. J., Hatfull, G. F. & Bloom, B. R. (1993). Rapid assessment of drug susceptibilities of *Mycobacterium tuberculosis* by means of luciferase reporter phages. *Science*, **260**, 819–822.
- Jones, W. D., Jr & David, H. L. (1970). Biosynthesis of a lipase by *Mycobacterium smegmatis* ATCC 607 infected by mycobacteriophage D29. *Am. Rev. Resp. Dis.* **102**, 818–820.
- Jones, W. D., Jr & David, H. L. (1971). Inhibition by rifampin of mycobacteriophage D29 replication in its drug-resistant host, *Mycobacterium smegmatis* ATCC 607. *Am. Rev. Resp. Dis.* **103**, 618–624.
- Lazraq, R., Moniz-Pereira, J., Clavel-Séres, S., Clément, F. & David, H. L. (1989). Restriction map of Mycobacteriophage D29 and its deletion mutant F5. *Acta Leprologica*, **7**, 234–238.
- Lee, M. H., Pascopella, L., Jacobs, W. R., Jr & Hatfull, G. F. (1991). Site-specific integration of mycobacteriophage L5: Integration-proficient vectors for *Mycobacterium smegmatis*, *Mycobacterium tuberculosis*, and bacille Calmette Guérin. *Proc. Natl Acad. Sci. USA*, **77**, 3220–3224.
- LeGendre, N. & Matsudaira, P. T. (1989). *A Practical Guide to Protein and Peptide Purification for Microsequencing* (Matsudaira, P. T., ed.), pp. 49–57, Academic Press, New York.
- Levin, M. E., Hendrix, R. W. & Casjens, S. R. (1993). A programmed translational frameshift is required for the synthesis of a bacteriophage lambda tail assembly protein. *J. Mol. Biol.* **234**, 124–139.
- McFadden, J. (1996). Recombination in mycobacteria. *Mol. Microbiol.* **21**, 205–211.
- Nesbit, C. E., Levin, M. E., Donnelly-Wu, M. K. & Hatfull, G. F. (1995). Transcriptional regulation of repressor synthesis in mycobacteriophage L5. *Mol. Microbiol.* **17**, 1045–1056.
- Oysaki, M. & Hatfull, G. F. (1992). The cohesive ends of mycobacteriophage L5 DNA. *Nucl. Acids Res.* **20**, 3251.
- Pascopella, L., Collins, F. M., Martin, J. M., Lee, M. H., Hatfull, G. F., Stover, C. K., Bloom, B. R. & Jacobs, W. R., Jr (1994). Use of *in vivo* complementation in *Mycobacterium tuberculosis* to identify a genomic fragment associated with virulence. *Infect. Immun.* **62**, 1313–1319.
- Pearson, R. E., Jurgensen, S., Sarkis, G. J., Hatfull, G. F. & Jacobs, W. R., Jr (1996). Construction of D29 shuttle plasmids and luciferase reporter phages for detection of mycobacteria. *Gene*, **183**, 129–136.
- Pelletier, I., Altenbuchner, J. & Mattes, R. (1995). A catalytic triad is required by the non-heme haloperoxidases to perform halogenation. *Biochim. Biophys. Acta*, **1250**, 149–157.
- Peña, C. E. A., Lee, M. H., Pedulla, M. L. & Hatfull, G. F. (1997). Characterization of the mycobacteriophage L5 attachment site, *attP*. *J. Mol. Biol.* **266**, 76–92.
- Popa, M. P., McKelvey, T. A., Hempel, J. & Hendrix, R. W. (1991). Bacteriophage HK97 structure: wholesale covalent cross-linking between the major head shell subunits. *J. Virol.* **65**, 3227–3237.
- Rennell, D. & Poteete, A. (1985). Phage P22 lysis genes: Nucleotide sequence and functional relationships with T4 and lambda genes. *Virology*, **143**, 280–289.
- Russell, R. L., Jann, G. J. & Froman, S. (1963). Lysogeny in the mycobacteria II. Some phage-host relationships of lysogenic mycobacteria. *Am. Rev. Resp. Dis.* **88**, 528–538.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Sarkis, G., Jacobs, W. R., Jr & Hatfull, G. F. (1995). L5 luciferase reporter mycobacteriophages: a sensitive tool for the detection and assay of live mycobacteria. *Mol. Microbiol.* **15**, 1055–1067.
- Shafer, R., Huber, U. & Franklin, R. M. (1977). Chemical and physical properties of mycobacteriophage D29. *Eur. J. Biochem.* **73**, 239–246.
- Snapper, S., Lugosi, L., Jekkel, A., Melton, R., Keiser, T., Bloom, B. R. & Jacobs, W. R., Jr (1988). Lysogeny and transformation in mycobacteria: stable expression of foreign genes. *Proc. Natl Acad. Sci. USA*, **85**, 6987–6991.
- Staden, R. (1986). The current status and portability of our sequence handling software. *Nucl. Acids Res.* **14**, 217–231.
- Stover, C. K., de la Cruz, V. F., Fuerst, T. R., Burlein, J. E., Benson, L. A., Bennett, L. T., Bansal, G. P., Young, J. F., Lee, M. H., Hatfull, G. F., Snapper, S., Barletta, R. G., Jacobs, W. R., Jr & Bloom, B. R. (1991). New use of BCG for recombinant vaccines. *Nature*, **351**, 456–460.
- Tokunaga, T. & Sellers, M. I. (1970). Changes in *Mycobacterium smegmatis* induced by lysogenization of L1 phage. In *Host-Virus Relationships in Mycobacterium, Nocardia and Actinomyces* (Juhasz, S. E. & Plummer, G., eds), pp. 119–133, Charles C. Thomas, Springfield, IL.
- Wolfframm, C., Lingens, F., Mutzel, R. & van Pee, K. (1993). Chloroperoxidase-encoding gene from *Pseudomonas pyrrocinia*: sequence, expression in heterologous hosts, and purification of the enzyme. *Gene*, **130**, 131–135.

Edited by J. Karn

(Received 28 October 1997; accepted 22 December 1997)